CENGAGE Learning®

10a. edición

ANÁLISIS NUMÉRICO

Richard L. Burden · Douglas J. Faires · Annette M. Burden

CENGAGE
Learning®

Análisis numérico,
10 a. ed.
Richard L. Burden, J. Douglas Faires y
Annette M. Burden

Director Editorial para Latinoamérica:
Ricardo H. Rodríguez

Editora de Adquisiciones para Latinoamérica:
Claudia C. Garay Castro

Gerente de Manufactura para Latinoamérica:
Antonio Mateos Martínez

Gerente Editorial de Contenidos en Español:
Pilar Hernández Santamarina

Gerente de Proyectos Especiales:
Luciana Rabuffetti

Coordinador de Manufactura:
Rafael Pérez González

Editora:
Ivonne Arciniega Torres

Diseño de portada:
Anneli Daniela Torres Arroyo

Imagen de portada:
© theromb/Shutterstock.com

Composición tipográfica:
Tsuki Marketing S.A. de C.V.
Gerardo Larios García

Traducido del libro *Numerical Analysis,* Tenth Edition
Richard L. Burden, J. Douglas Faires, Annette M. Burden
Publicado en inglés por Cengage Learning
© 2016, 2011, 2005
ISBN: 978-1-305-25366-7

Datos para catalogación bibliográfica:
Burden, Faires y Burden
Análisis numérico, 10a. ed.
ISBN: 978-607-526-411-0

Visite nuestro sitio en:
http://latinoamerica.cengage.com

Impreso en México
1 2 3 4 5 6 7 20 19 18 17

Análisis numérico

DÉCIMA EDICIÓN

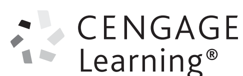
Richard L. Burden
Youngstown University

J. Douglas Faires
Youngstown University

Annette M. Burden
Youngstown University

Traducción:
Mara Paulina Suárez Moreno
Traductora profesional

Revisión técnica:
Wilmar Alberto Díaz Ossa
Mágister en matemáticas aplicadas
Profesor en la Universidad Distrital Francisco José de Caldas



Australia • Brasil • Corea • España • Estados Unidos • Japón • México • Reino Unido • Singapur

Contenido

Material en línea

El siguiente material se encuentra disponible en línea:

- Conjuntos de ejercicios
- Preguntas de análisis
- Conceptos clave
- Revisión de capítulo
- Bibliografía
- Respuestas a ejercicios seleccionados
- Índice
- Índice de algoritmos
- Glosario de notación
- Trigonometría
- Gráficas comunes

Ingrese a **www.cengage.com**, busque el libro por el ISBN e ingrese el siguiente código de acceso:

Prefacio



Acerca del texto

Este texto se escribió para una secuencia de cursos sobre la teoría y la aplicación de técnicas de aproximación numérica. Está diseñado principalmente para los estudiantes avanzados de matemáticas y ciencias e ingeniería de nivel básico que han terminado por lo menos el primer año de la sucesión de cálculo universitario estándar. La familiaridad con los fundamentos de álgebra de matrices y las ecuaciones diferenciales es útil, pero existe suficiente material introductorio sobre estos temas, por lo que los cursos relacionados con ellos no son por fuerza requisitos previos.

Se han usado ediciones previas de *Análisis numérico* en distintas situaciones. En algunos casos, el análisis matemático subyacente al desarrollo de técnicas de aproximación recibió más énfasis que los métodos; en otros, el énfasis era inverso. El libro se ha utilizado como referencia central para los cursos de nivel posgrado en ingeniería, matemáticas, programas de ciencias de la computación y cursos de primer año sobre análisis introductorio impartidos en universidades internacionales. Hemos adaptado el libro para ajustarnos a los diferentes usuarios sin comprometer nuestro objetivo original:

> *Presentar técnicas modernas de aproximación; explicar cómo, porqué y cuándo se puede esperar que funcionen, y proporcionar las bases para más estudios de análisis numérico y cálculo científico*.

El libro contiene material suficiente para al menos un año completo de estudio, pero esperamos que muchas personas lo usen solamente para un curso de un solo plazo. En dichos cursos, los estudiantes aprenden a identificar los tipos de problemas que requieren técnicas numéricas para su solución y observan ejemplos de error de propagación que se pueden presentar al aplicar métodos numéricos. Ellos se aproximan con precisión a la solución de los problemas que no se pueden resolver exactamente y aprenden técnicas comunes para calcular las cotas del error para sus aproximaciones. El resto del texto sirve como referencia para los métodos que no se consideran en el curso. El tratamiento tanto para los cursos de todo un año como para los cursos de un solo plazo es consistente con la filosofía del texto.

Prácticamente todos los conceptos en el texto están ilustrados con un ejemplo y esta edición contiene más de 2 500 ejercicios probados en clase que van desde aplicaciones fundamentales de métodos y algoritmos hasta generalizaciones y extensiones de la teoría. Además, los conjuntos de ejercicios incluyen varios problemas aplicados de diversas áreas de la ingeniería, así como de la física, la informática, la biología y las ciencias económicas y sociales. Las aplicaciones, seleccionadas de forma clara y concisa, demuestran la manera en la que las técnicas numéricas pueden y, a menudo, se aplican en situaciones de la vida real.

Se ha desarrollado una serie de paquetes de software, conocidos como Sistemas de Álgebra Computacional (CAS) para producir cálculos matemáticos simbólicos. Maple©, Mathematica© y MATLAB© destacan en el ambiente académico. Existen versiones para estudiantes de estos paquetes de software a un precio razonable para la mayoría de los

sistemas informáticos. Además, ahora existe Sage, un sistema de código abierto. La información sobre este sistema se puede encontrar en el sitio

http://www.sagemath.org

Aunque existen diferencias entre los paquetes, tanto de desempeño como de precio, todos efectúan operaciones de cálculo y algebraicas estándar.

Los resultados en muchos de nuestros ejemplos y ejercicios se han generado por medio de problemas para los que se *pueden* determinar valores exactos ya que esto permite mejorar el desempeño de la supervisión del método de aproximación. Además, para muchas técnicas numéricas, el análisis de error requiere limitar una derivada ordinaria o parcial superior de una función, lo cual puede ser una tarea tediosa y que no es especialmente instructiva una vez que se dominan las técnicas de cálculo. Por lo tanto, tener un paquete informático simbólico puede ser muy útil en el estudio de las técnicas de aproximación porque, a menudo, las soluciones exactas se obtienen fácilmente por medio de cálculos simbólicos. Las derivadas se pueden obtener simbólicamente de manera rápida y, con frecuencia, un poco de perspectiva permite que el cálculo simbólico también ayude al proceso de limitación.

Algoritmos y programas

En nuestra primera edición, presentamos una característica que en ese tiempo era innovadora y controvertida. En lugar de presentar nuestras técnicas de aproximación en un lenguaje de programación específico (FORTRAN dominaba en esa época), dimos algoritmos en un pseudocódigo que conduciría a un programa bien estructurado en una variedad de lenguajes. Al inicio de la segunda edición, listamos los programas en lenguajes específicos en el *Manual del Instructor* para el libro y el número de estos lenguajes se ha incrementado en las ediciones posteriores. Ahora, codificamos los programas y están disponibles en línea en varios de los lenguajes de programación más comunes y hojas de cálculo CAS. Todos estos se encuentran en el sitio web que acompaña al libro (consulte “Complementos” [disponibles sólo para la edición en inglés y se venden por separado]).

Para cada algoritmo existe un programa escrito en Fortran, Pascal, C y Java. Además, codificamos los programas usando Maple, Mathematica y MATLAB. Esto debería garantizar que haya un conjunto de programas disponibles para la mayoría de los sistemas de computadora comunes.

Todos los programas se ilustran con un programa muestra, correlacionado estrechamente con el texto. Esto permite ejecutar inicialmente el programa en el lenguaje de su elección para observar el formato de entrada y salida. A continuación, los programas se pueden modificar para otros problemas al realizar cambios menores. Los formatos de entrada y salida son, en la medida de lo posible, iguales en cada uno de los sistemas de programación. Esto permite al instructor usar los programas para analizarlos de manera genérica, de manera independiente del sistema de programación particular que use un solo estudiante.

Los programas están diseñados para ejecutarse en una computadora con configuraciones mínimas y se proporciona en formato ASCII para permitir flexibilidad de uso. Esto hace modificarlos mediante cualquier editor o procesador de palabras que cree archivos ASCII estándar. (En general, también reciben el nombre de archivos “sólo texto”.) Se incluyen archivos README extensos junto con los archivos del programa, por lo que las peculiaridades de diferentes sistemas de programación se pueden abordar de manera individual. Los archivos README se presentan tanto en formato ASCII como en archivos PDF. Puesto que se desarrolla software nuevo, los programas se actualizarán y colocarán en el sitio web del libro.

Para la mayoría de los sistemas de programación se necesita software adecuado, como un compilador para Pascal, Fortran y C, o uno de los Sistemas de Álgebra Computacional (Maple, Mathematica y MATLAB). Las implementaciones Java son una excepción. Usted necesita el sistema para ejecutar los programas, pero Java se puede descargar gratis desde diferentes sitios web. La mejor forma de obtener Java es utilizar un motor de búsqueda para buscar el nombre, seleccionar el sitio web de descarga y seguir las instrucciones provistas en el mismo.

Nuevo en esta edición

La primera edición de este libro se publicó hace más de 35 años, en la década anterior a los avances más importantes en técnicas numéricas, los cuales reflejan la amplia disponibilidad del equipo informático. En nuestras revisiones del libro, hemos añadido técnicas nuevas, en un intento por mantener nuestro trato actual. Para continuar con esta tendencia, realizamos una serie de cambios significativos en esta edición:

- Reescribimos algunos de los ejemplos en el libro para enfatizar mejor el problema que se va a resolver antes de proporcionar la solución. Agregamos pasos adicionales a algunos de los ejemplos para mostrar explícitamente los cálculos requeridos para los primeros pasos de los procesos de iteración. Esto da a los lectores una forma de probar y depurar los programas que escriben para problemas similares a los ejemplos.
- Dividimos los ejercicios del capítulo en computacionales, aplicados y teóricos para proporcionar más flexibilidad al instructor al asignar la tarea. Casi en todas las situaciones computacionales los ejercicios se emparejaron de forma par e impar. Puesto que los problemas impares se resuelven en la última parte del texto, si los problemas pares se asignan como tarea los estudiantes podrán trabajar los problemas impares y verificar sus respuestas antes de resolver el problema par.
- Agregamos muchos ejercicios aplicados nuevos al texto.
- Incluimos preguntas de análisis después de cada sección del capítulo, principalmente para uso del instructor en los cursos en línea.
- Renombramos la última sección de cada capítulo y la dividimos en cuatro subsecciones: Software numérico, Preguntas de análisis, Conceptos clave y Revisión del capítulo (que se encuentran disponibles en línea). La mayoría de las Preguntas de análisis llevan al estudiante hacia áreas modernas de investigación en el desarrollo de software.
- Reorganizamos partes del texto para facilitar la instrucción en línea.
- Agregamos diapositivas en PowerPoint para complementar el material de lectura (disponibles sólo para la versión en inglés).
- Actualizamos el material bibliográfico para reflejar las nuevas ediciones de los libros que consultamos. Agregamos nuevas fuentes que antes no estaban disponibles.

Como siempre con nuestras revisiones, examinamos todos los enunciados para determinar si estaban redactados de la mejor forma relacionada con lo que tratamos de describir.

Complementos

Los autores crearon un sitio web que acompaña el libro, el cual contiene los materiales complementarios que se mencionan más adelante. El sitio web se encuentra en

https://sites.google.com/site/numericalanalysis1burden/

es para estudiantes e instructores. Parte del material en el sitio web es sólo para uso del instructor. Los instructores pueden acceder al material protegido al contactar a los autores para obtener la contraseña **(disponibles sólo para la versión en inglés y se venden por separado)**.

Algunos de los complementos también se pueden obtener en

https://www.cengagebrain.com

mediante la búsqueda del ISBN **(disponibles sólo para la versión en inglés)**.

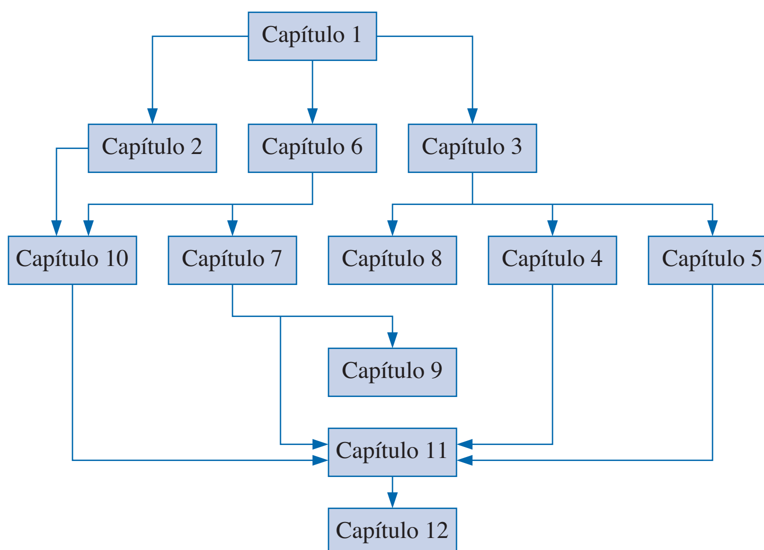
1. *Ejemplos del programa para estudiantes* que contienen código Maple, Matlab y Excel para que los estudiantes lo utilicen para resolver problemas del texto. Está organizado en forma paralela al texto, capítulo por capítulo. Se ilustran los comandos en estos sistemas. Los comandos se presentan en segmentos de programa muy cortos para mostrar la forma de resolver los ejercicios sin programación extensa.
2. *Conferencias para estudiantes* que contienen una perspectiva adicional al contenido del capítulo. Estas conferencias se escribieron principalmente para el aprendiz en línea, pero pueden ser útiles para estudiantes que toman el curso de manera tradicional.
3. *Guía de estudio para el estudiante* que contiene las soluciones de muchos de los problemas. Los primeros dos capítulos de esta guía están disponibles en el sitio web del libro en formato PDF, por lo que los posibles usuarios pueden decir si los encuentran suficientemente útiles. Toda la guía se puede obtener sólo a partir del editor al llamar a Cengage Learning Customer & Sales Support al 1-800-354-9706 o al ordenar en línea en http://www.cengagebrain.com/.
4. *Programas de algoritmo* que son programas completos escritos en Maple, Matlab, Mathematica, C, Pascal, Fortran y Java para todos los algoritmos en el texto. Estos programas están previstos para estudiantes más experimentados en lenguajes de programación.
5. *Diapositivas en PowerPoint para el instructor* en formato PDF para uso del instructor, tanto para cursos tradicionales como en línea. Contacte a los autores para obtener la contraseña.
6. *Manual del instructor* que provee respuestas y soluciones para todos los ejercicios en el libro. Los resultados de los cálculos en el Manual del Instructor se regeneraron para esta edición mediante los programas en el sitio web para garantizar la compatibilidad entre los diferentes sistemas de programación. Contacte a los autores para obtener la contraseña.
7. *Pruebas muestra para el instructor* para uso del instructor. Contacte a los autores para obtener la contraseña.
8. *Erratas*.

Posibles sugerencias para el curso

Análisis numérico está diseñado para proporcionar flexibilidad a los instructores al elegir los temas, así como en el nivel de rigor teórico y en el énfasis sobre las aplicaciones. Junto con estos objetivos, proporcionamos referencias detalladas para los resultados que no se demuestran en el texto y para las aplicaciones que se utilizan para señalar la importancia práctica de los métodos. Las referencias de los textos citados son aquellas que se pueden encontrar con mayor facilidad en las bibliotecas universitarias y se actualizaron para reflejar ediciones recientes. También incluimos citas de los documentos originales de investigación cuando sentimos que este material es accesible para nuestra audiencia prevista. Todos los materiales consultados se han indexado en las ubicaciones adecuadas en el texto y se incluye la información sobre la convocatoria de la Biblioteca del Congreso para el material de referencia para permitir localizarlo fácilmente al buscarlo en la biblioteca.

El siguiente diagrama de flujo indica los requisitos previos del capítulo. La mayor parte de las secuencias posibles que se pueden generar a partir de este diagrama fueron enseñadas por los autores en Youngstown State University.

El material en esta edición debe permitir a los instructores preparar un curso universitario de álgebra lineal numérica para los estudiantes que no han estudiado análisis numérico antes. Esto se puede realizar al cubrir los capítulos 1, 6, 7 y 9.



Reconocimientos

Afortunadamente obtuvimos las impresiones de muchos de nuestros estudiantes y colegas sobre ediciones anteriores de este libro y hemos tomado estos comentarios muy en serio. Hemos intentado incluir todas las sugerencias que complementan la filosofía del libro y estamos sumamente agradecidos con todos aquellos que se han tomado el tiempo para contactarnos y proporcionarnos formas de mejorar las versiones posteriores.

Nos gustaría agradecer especialmente a las siguientes personas, cuyas sugerencias han sido útiles para ésta y otras ediciones previas.

Douglas Carter,

John Carroll, Dublin University

Yavuz Duman, T.C. Istanbul Kultur Universitesi

Neil Goldman,

Christopher Harrison,

Teryn Jones, Youngstown State University

Aleksandar Samardzic, University of Belgrade

Mikhail M. Shvartsman, University of St. Thomas

Dennis C. Smolarski, Santa Clara University

Dale Smith, Comcast

Nos gustaría agradecer a la doctora Barbara T. Faires por su cooperación al proporcionarnos los materiales necesarios para hacer posible esta revisión. Su gentileza durante estos tiempos difíciles fue muy apreciada.

Siguiendo nuestra práctica en las ediciones pasadas, hemos obtenido la ayuda de los estudiantes de la Youngstown State University. Nuestro capaz asistente para esta edición fue Teryn Jones, quien trabajó en los applets de Java. Nos gustaría agradecer a Edward R. Burden, un estudiante de doctorado en Ingeniería eléctrica en la Ohio State University, quien ha estado verificando todos los problemas de aplicación y el material nuevo en el texto. También nos gustaría expresar nuestra gratitud con nuestros colegas en la facultad y administración de la Youngstown State University por darnos la oportunidad y las instalaciones para completar este proyecto.

Nos gustaría agradecer a algunas de las personas que hicieron contribuciones importantes a la historia de los métodos numéricos. Herman H. Goldstine escribió un libro excelente titulado *A History of Numerical Analysis from the 16th Through the 19th Century* [Golds]. Otra fuente de excelente conocimiento matemático histórico es el archivo MacTutor History of Mathematics en la University of St. Andrews en Escocia. Fue creado por John J. O'Connor y Edmund F. Robertson y tiene la dirección de internet

http://www-gap.dcs.st-and.ac.uk/∼history/

Se ha dedicado una cantidad increíble de trabajo a la creación de material en este sitio web y descubrimos que la información es invariablemente precisa. Finalmente, gracias a todos los que contribuyen con Wikipedia, quienes han agregado su experiencia a ese sitio web para que otros puedan beneficiarse de su conocimiento.

Por último, gracias nuevamente a aquellos que han dedicado tiempo y esfuerzo en contactarnos durante años. Ha sido grandioso escuchar a tantos estudiantes y docentes, quienes utilizan nuestro libro para su primera exposición al estudio de métodos numéricos. Esperamos que esta edición continúe con este intercambio y se sume al deleite de los estudiantes de análisis numérico. Si tiene alguna sugerencia para mejorar ediciones futuras del libro, como siempre, agradeceremos sus comentarios. Contáctenos fácilmente por correo electrónico a través de las direcciones enumeradas más abajo.

Richard L. Burden
rlburden@ysu.edu

Annette M. Burden
amburden@ysu.edu

Esta edición está dedicada a la memoria de
J. Douglas Faires.
Doug fue amigo, colega y coautor durante más de 40 años.
Se le echará mucho de menos.

CAPÍTULO

1

Preliminares matemáticos y análisis de error

Introducción

Al comenzar los cursos de química, estudiamos la *ley del gas ideal*,

$$PV = NRT,$$

que relaciona la presión P, el volumen V, la temperatura T y el número de moles N de un gas "ideal". En esta ecuación, R es una contante que depende del sistema de medición.

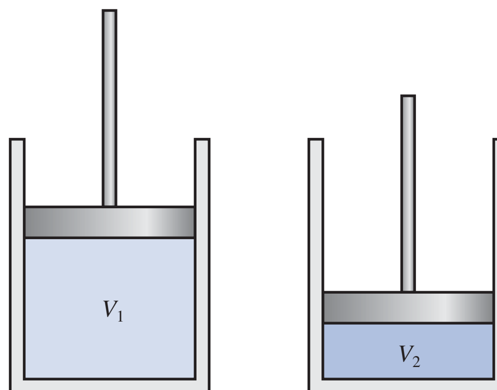
Suponga que se realizan dos experimentos para evaluar esta ley, mediante el mismo gas en cada caso. En el primer experimento,

$$P = 1.00 \text{ atm}, \qquad V = 0.100 \text{ m}^3,$$
$$N = 0.00420 \text{ mol}, \qquad R = 0.08206.$$

La ley del gas ideal predice que la temperatura del gas es

$$T = \frac{PV}{NR} = \frac{(1.00)(0.100)}{(0.00420)(0.08206)} = 290.15 \text{ K} = 17°\text{C}.$$

Sin embargo, cuando medimos la temperatura del gas, encontramos que la verdadera temperatura es 15°C.



A continuación, repetimos el experimento utilizando los mismos valores de R y N, pero incrementamos la presión en un factor de dos y reducimos el volumen en ese mismo factor. El producto PV sigue siendo el mismo, por lo que la temperatura prevista sigue siendo 17°C. Sin embargo, ahora encontramos que la temperatura real del gas es 19°C.

Claramente, se sospecha la ley de gas ideal, pero antes de concluir que la ley es inválida en esta situación, deberíamos examinar los datos para observar si el error se puede atribuir a los resultados del experimento. En este caso, podríamos determinar qué tan precisos deberían ser nuestros resultados experimentales para evitar que se presente un error de esta magnitud.

El análisis del error involucrado en los cálculos es un tema importante en análisis numérico y se presenta en la sección 1.2. Esta aplicación particular se considera en el ejercicio 26 de esa sección.

Este capítulo contiene una revisión breve de los temas del cálculo de una sola variable que se necesitarán en capítulos posteriores. Un conocimiento sólido de cálculo es fundamental para comprender el análisis de las técnicas numéricas y sería preciso efectuar una revisión más rigurosa para quienes no han estado en contacto con este tema durante un tiempo. Además, existe una introducción a la convergencia, el análisis de error, la representación de números en lenguaje de máquina y algunas técnicas para clasificar y minimizar el error computacional.

1.1 Revisión de cálculo

Límites y continuidad

Los conceptos de *límite* y *continuidad* de una función son fundamentales para el estudio del cálculo y constituyen la base para el análisis de las técnicas numéricas.

Definición 1.1 Una función f definida en un conjunto X de números reales que tiene el **límite** L a x_0, escrita como

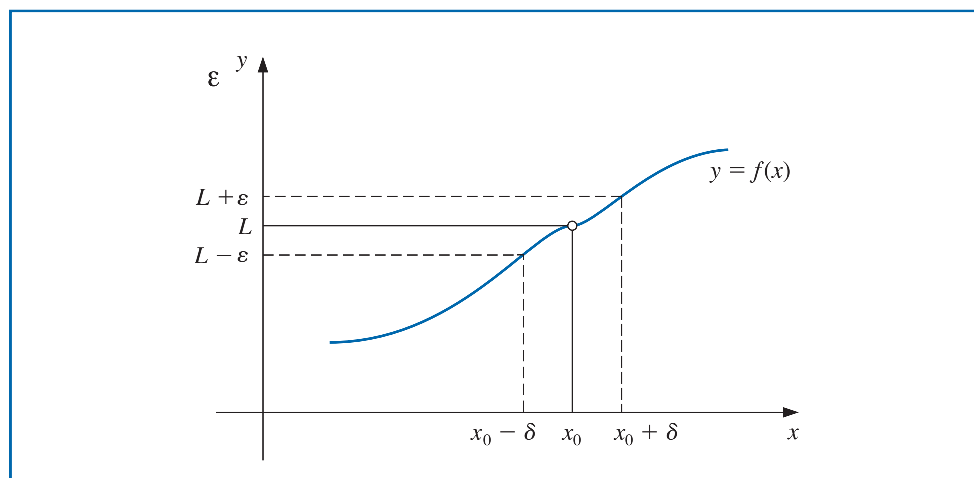
$$\lim_{x \to x_0} f(x) = L,$$

si, dado cualquier número real $\varepsilon > 0$, existe un número real $\delta > 0$, de tal forma que

$$|f(x) - L| < \varepsilon, \qquad \text{siempre que} \quad x \in X \quad \text{y} \quad 0 < |x - x_0| < \delta.$$

(consulte la figura 1.1) ■

Figura 1.1



Definición 1.2 Sea f una función definida en un conjunto X de números reales y $x_0 \in X$. Entonces f es **continua** en x_0 si

$$\lim_{x \to x_0} f(x) = f(x_0).$$

La función f es **continua en el conjunto** X si es continua en cada número en X. ■

Los conceptos básicos de cálculo y sus aplicaciones se desarrollaron a finales del siglo XVII y a principios del XVIII, pero los conceptos matemáticamente precisos de límites y continuidad se describieron hasta la época de Augustin Louis Cauchy (1789–1857), Heinrich Eduard Heine (1821–1881) y Karl Weierstrass (1815–1897) a finales del siglo XIX.

El conjunto de todas las funciones que son continuas en el conjunto X se denota como $C(X)$. Cuando X es un intervalo de la recta real, se omiten los paréntesis en esta notación. Por ejemplo, el conjunto de todas las funciones continuas en el intervalo cerrado $[a, b]$ se denota como $C[a, b]$. El símbolo $\mathbb{R}$ denota el conjunto de todos los números reales, que también tiene la notación del intervalo $(-\infty, \infty)$. Por eso el conjunto de todas las funciones que son continuas en cada número real se denota mediante $C(\mathbb{R})$ o mediante $C(-\infty, \infty)$.

El *límite de una sucesión* de números reales o complejos se define de manera similar.

Definición 1.3 Sea $\{x_n\}_{n=1}^{\infty}$ una sucesión infinita de números reales. Esta sucesión tiene el **límite** x (**converge a** x) si, para cualquier $\varepsilon > 0$, existe un entero positivo $N(\varepsilon)$ tal que $|x_n - x| < \varepsilon$ siempre que $n > N(\varepsilon)$. La notación

$$\lim_{n \to \infty} x_n = x, \quad \text{o} \quad x_n \to x \quad \text{en} \quad n \to \infty,$$

significa que la sucesión $\{x_n\}_{n=1}^{\infty}$ converge a x. ■

Teorema 1.4 Si f es una función definida en un conjunto X de números reales y $x_0 \in X$, entonces los siguientes enunciados son equivalentes:

a. f es continua en x_0;

b. Si $\{x_n\}_{n=1}^{\infty}$ es cualquier sucesión en X, que converge a x_0, entonces $\lim_{n \to \infty} f(x_n) = f(x_0)$. ■

Se asumirá que las funciones que consideraremos al analizar los métodos numéricos son continuas porque éste es el requisito mínimo para una conducta predecible. Las funciones que no son continuas pueden pasar por alto puntos de interés, lo cual puede causar dificultades al intentar aproximar la solución de un problema.

Diferenciabilidad

Las suposiciones más sofisticadas sobre una función por lo general conducen a mejores resultados de aproximación. Por ejemplo, normalmente una función con una gráfica suave se comportaría de forma más predecible que una con numerosas características irregulares. La condición de uniformidad depende del concepto de la derivada.

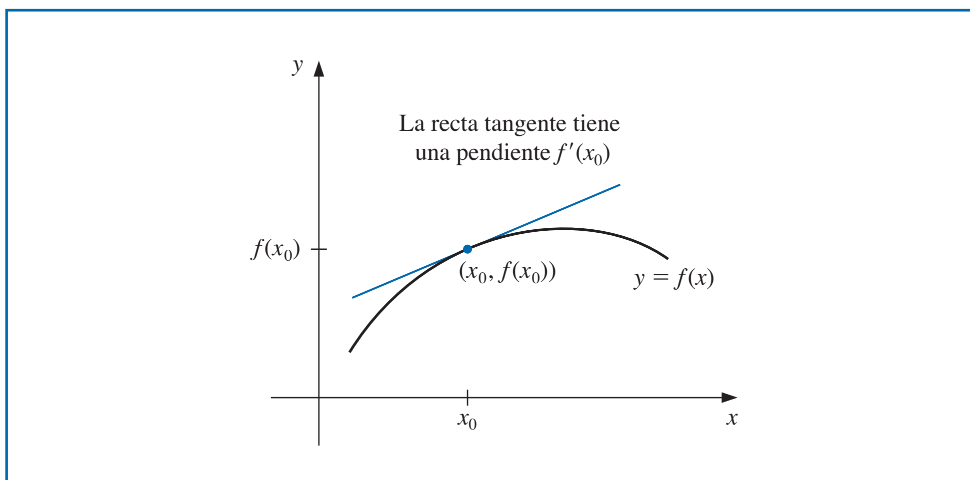
Definición 1.5 Si f es una función definida en un intervalo abierto que contiene x_0. La función f es **diferenciable** en x_0 si

$$f'(x_0) = \lim_{x \to x_0} \frac{f(x) - f(x_0)}{x - x_0}$$

existe. El número $f'(x_0)$ recibe el nombre de **derivada** de f en x_0. Una función que tiene una derivada en cada número en un conjunto X es **diferenciable en** X. ■

La derivada de f en x_0 es la pendiente de la recta tangente a la gráfica de f en $(x_0, f(x_0))$, como se muestra en la figura 1.2.

Figura 1.2



Teorema 1.6 Si la función f es diferenciable en x_0, entonces f es continua en x_0. ■

El teorema atribuido a Michel Rolle (1652–1719) apareció en 1691 en un tratado poco conocido titulado *Méthode pour résoundre les égalites (Método para resolver las igualdades)*. Originalmente, Rolle criticaba el cálculo desarrollado por Isaac Newton y Gottfried Leibniz, pero después se convirtió en uno de sus defensores.

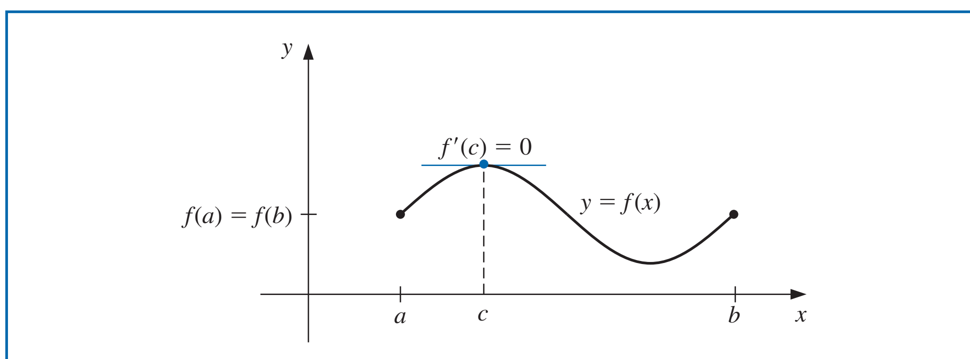
Los siguientes teoremas son de importancia fundamental al deducir los métodos para estimación del cálculo de error. Las pruebas de estos teoremas y los otros resultados sin referencias en esta sección se pueden encontrar en cualquier texto de cálculo estándar.

El conjunto de todas las funciones que tienen derivadas continuas n en X se denota como $C^n(X)$ y el conjunto de funciones que tienen derivadas de todos los órdenes en X se denota como $C^\infty(X)$. Las funciones polinomial, racional, trigonométrica, exponencial y logarítmica se encuentran en $C^\infty(X)$, donde X consiste en todos los números para los que se definen las funciones. Cuando X es un intervalo de la recta real, de nuevo se omiten los paréntesis en esta notación.

Teorema 1.7 **(Teorema de Rolle)**

Suponga que $f \in C[a, b]$ y f es diferenciable en (a, b). Si $f(a) = f(b)$, entonces existe un número c en (a, b) con $f'(c) = 0$. (Consulte la figura 1.3.) ■

Figura 1.3

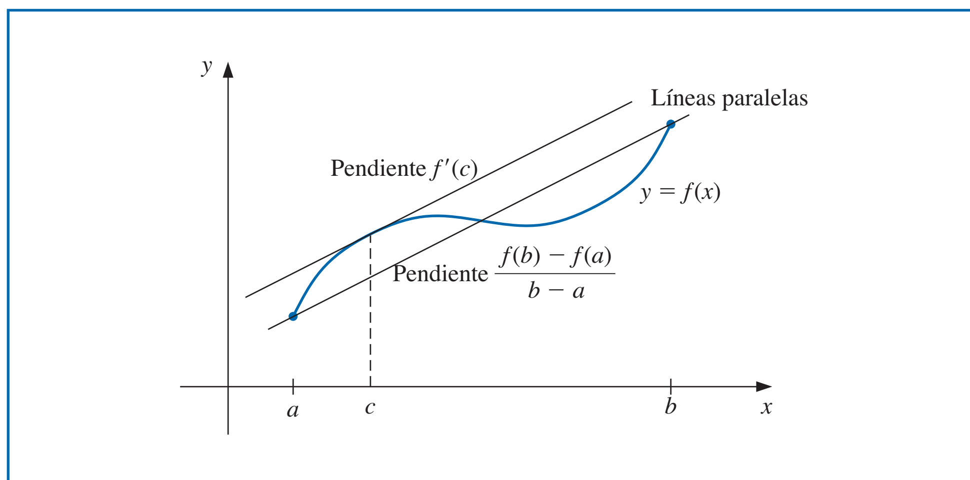


Teorema 1.8 **(Teorema del valor medio)**

Si $f \in C[a, b]$ y f es diferenciable en (a, b), entonces existe un número c en (a, b) con (consulte la figura 1.4.)

$$f'(c) = \frac{f(b) - f(a)}{b - a}.$$ ■

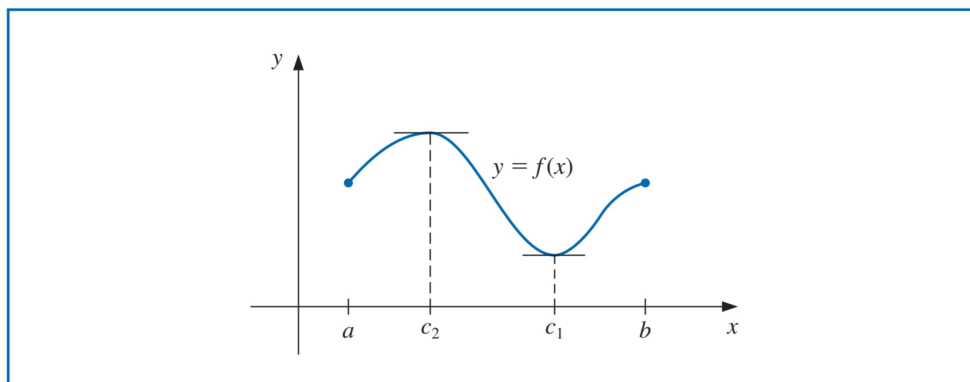
Figura 1.4



Teorema 1.9 **(Teorema del valor extremo)**

Si $f \in C[a, b]$, entonces existe $c_1, c_2 \in [a, b]$ con $f(c_1) \leq f(x) \leq f(c_2)$, para todas las $x \in [a, b]$. Además, si f es diferenciable en (a, b), entonces se presentan los números c_1 y c_2 ya sea en los extremos de $[a, b]$ o donde f' es cero. (Consulte la figura 1.5.) ■

Figura 1.5



Ejemplo 1 Encuentre los valores mínimo absoluto y máximo absoluto de

$$f(x) = 2 - e^x + 2x$$

en los intervalos **a**) [0, 1] y **b**) [1, 2].

Solución Comenzamos por derivar $f(x)$ para obtener

$$f'(x) = -e^x + 2.$$

$f'(x) = 0$ cuando $-e^x + 2 = 0$ o de forma equivalente, cuando $e^x = 2$. Al tomar el logaritmo natural de ambos lados de la ecuación obtenemos

$$\ln (e^x) = \ln (2) \text{ o } x = \ln (2) \approx 0.69314718056$$

a) Cuando el intervalo es [0, 1], el extremo absoluto debe ocurrir en $f(0)$, $f(\ln(2))$, o $f(1)$. Al evaluar, tenemos

$$f(0) = 2 - e^0 + 2(0) = 1$$
$$f(\ln(2)) = 2 - e^{\ln(2)} + 2\ln(2) = 2\ln(2) \approx 1.38629436112$$
$$f(1) = 2 - e + 2(1) = 4 - e \approx 1.28171817154.$$

Por lo tanto, el mínimo absoluto de $f(x)$ en [0, 1] es $f(0) = 1$ y el máximo absoluto es $f(\ln(2)) = 2\ln(2)$.

b) Cuando el intervalo es [1, 2], sabemos que $f'(x) \neq 0$, por lo que el extremo absoluto se presenta en $f(1)$ y $f(2)$. Por lo tanto, $f(2) = 2 - e^2 + 2(2) = 6 - e^2 \approx -1.3890560983$.
El mínimo absoluto en [1, 2] es $6 - e^2$ y el máximo absoluto es 1.
Observamos que

$$\max_{0\leq x\leq 2} |f(x)| = |6 - e^2| \approx 1.3890560983.$$ ■

En general, el siguiente teorema no se presenta en un curso de cálculo básico, pero se deriva al aplicar el teorema de Rolle sucesivamente a f, f', . . . , y finalmente, a $f^{(n-1)}$. Este resultado se considera en el ejercicio 26.

Teorema 1.10 (Teorema generalizado de Rolle)

Suponga que $f \in C[a, b]$ es n veces diferenciable en (a, b). Si $f(x) = 0$ en los $n + 1$ números distintos a $a \leq x_0 < x_1 < \ldots < x_n \leq b$, entonces un número c en (x_0, x_n) y, por lo tanto, en (a, b) existe con $f^{(n)}(c) = 0$. ■

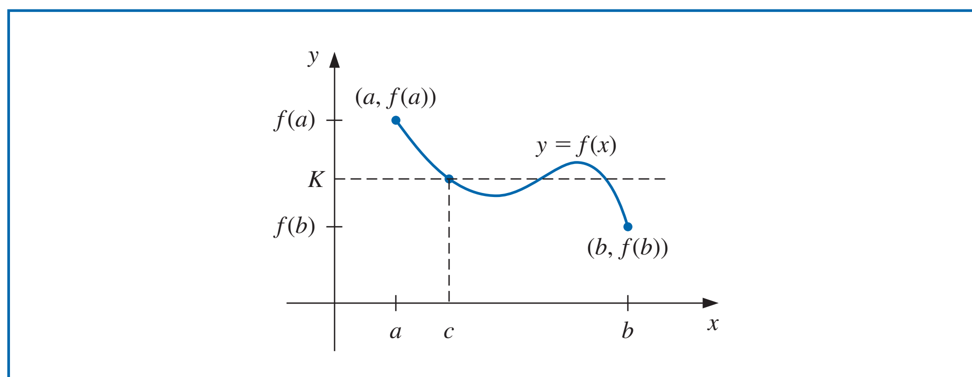
También utilizaremos con frecuencia el teorema del valor intermedio. A pesar de que esta declaración parece razonable, su prueba va más allá del alcance del curso habitual de cálculo. Sin embargo, se puede encontrar en muchos textos de análisis (consulte, por ejemplo, [Fu], p. 67).

Teorema 1.11 (Teorema del valor intermedio)

Si $f \in C[a, b]$ y K es cualquier número entre $f(a)$ y $f(b)$, entonces existe un número c en (a, b) para el cual $f(c) = K$. ■

La figura 1.6 muestra una opción para el número garantizada por el teorema del valor intermedio. En este ejemplo, existen otras dos posibilidades.

Figura 1.6



Ejemplo 2 Muestre que $x^5 - 2x^3 + 3x^2 - 1 = 0$ tiene una solución en el intervalo [0, 1].

Solución Considere la función definida por $f(x) = x^5 - 2x^3 + 3x^2 - 1$. La función f es continua en [0, 1]. Además,

$$f(0) = -1 < 0 \qquad \text{y} \qquad 0 < 1 = f(1).$$

Por lo tanto, el teorema del valor intermedio implica que existe un número c, con $0 < c < 1$, para el cual $c^5 - 2c^3 + 3c^2 - 1 = 0$. ■

Como se observa en el ejemplo 2, el teorema del valor intermedio se utiliza para determinar cuándo existen soluciones para ciertos problemas. Sin embargo, no provee un medio eficiente para encontrar estas soluciones. Este tema se considera en el capítulo 2.

Integración

El otro concepto básico del cálculo que se utilizará ampliamente es la integral de Riemann.

Definición 1.12 La **integral de Riemann** de la función f en el intervalo $[a, b]$ es el siguiente límite, siempre y cuando exista:

$$\int_a^b f(x)\,dx = \lim_{\text{máx } \Delta x_i \to 0} \sum_{i=1}^{n} f(z_i)\,\Delta x_i,$$

donde los números $x_0, x_1, \ldots, x_n$ satisfacen $a = x_0 \leq x_1 \leq \ldots \leq x_n = b$, donde $\Delta x_i = x_i - x_{i-1}$, para cada $i = 1, 2, \ldots, n$, y z_i se selecciona de manera arbitraria en el intervalo $[x_{i-1}, x_i]$. ■

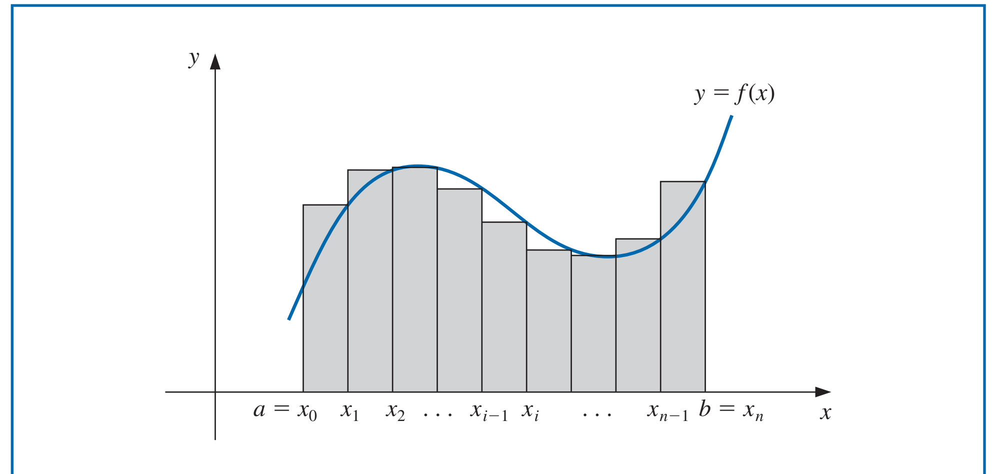
George Fredrich Berhard Riemann (1826–1866) realizó muchos de los descubrimientos importantes para clasificar las funciones que tienen integrales. También realizó trabajos fundamentales en geometría y la teoría de funciones complejas y se le considera uno de los matemáticos prolíferos del siglo XIX.

Una función f que es continua en un intervalo $[a, b]$ es también Riemann integrable en $[a, b]$. Esto nos permite elegir, para conveniencia computacional, los puntos x_i se separarán equitativamente en $[a, b]$ para cada $i = 1, 2, \ldots, n$, para seleccionar $z_i = x_i$. En este caso,

$$\int_a^b f(x)\,dx = \lim_{n \to \infty} \frac{b-a}{n} \sum_{i=1}^{n} f(x_i),$$

donde los números mostrados en la figura 1.7, como x_i, son $x_i = a + i(b - a)/n$.

Figura 1.7



Se necesitarán otros dos resultados en nuestro estudio para análisis numérico. El primero es una generalización del teorema del valor promedio para integrales.

Teorema 1.13 **(Teorema del valor promedio para integrales)**

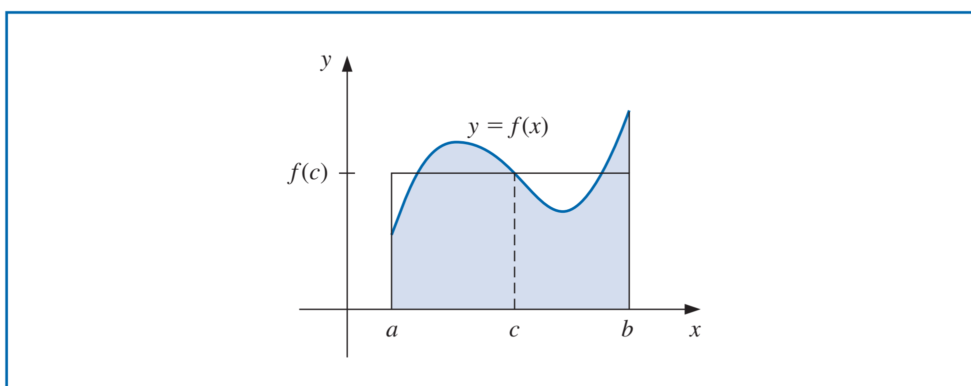
Suponga que $f \in C[a, b]$, la integral de Riemann de g existe en $[a, b]$, y $g(x)$ no cambia de signo en $[a, b]$. Entonces existe un número c en (a, b) con

$$\int_a^b f(x)g(x)\,dx = f(c)\int_a^b g(x)\,dx. \qquad \blacksquare$$

Cuando $g(x) \equiv 1$, el teorema 1.13 es el teorema del valor medio para integrales. Éste proporciona el **valor promedio** de la función f sobre el intervalo $[a, b]$ como (consulte la figura 1.8.)

$$f(c) = \frac{1}{b-a}\int_a^b f(x)\,dx.$$

Figura 1.8



En general la prueba del teorema 1.13 no se da en un curso básico de cálculo, pero se puede encontrar en muchos textos de análisis (consulte, por ejemplo, [Fu], p. 162)

Polinomios y series de Taylor

El teorema final en esta revisión de cálculo describe los polinomios de Taylor. Estos polinomios se usan ampliamente en el análisis numérico.

Teorema 1.14 **(Teorema de Taylor)**

Brook Taylor (1685–1731) describió esta serie en 1715 en el artículo *Methodus incrementorum directa et inversa (Métodos para incrementos directos e inversos)*. Isaac Newton, James Gregory y otros ya conocían algunos casos especiales del resultado y, probablemente, el resultado mismo.

Suponga que $f \in C^n[a, b]$, $f^{(n+1)}$ existe en $[a, b]$, y $x_0 \in [a, b]$. Para cada $x \in [a, b]$, existe un número $\xi(x)$ entre x_0 y x con

$$f(x) = P_n(x) + R_n(x),$$

donde

$$\begin{aligned} P_n(x) &= f(x_0) + f'(x_0)(x - x_0) + \frac{f''(x_0)}{2!}(x - x_0)^2 + \cdots + \frac{f^{(n)}(x_0)}{n!}(x - x_0)^n \\ &= \sum_{k=0}^{n} \frac{f^{(k)}(x_0)}{k!}(x - x_0)^k \end{aligned}$$

y

$$R_n(x) = \frac{f^{(n+1)}(\xi(x))}{(n+1)!}(x - x_0)^{n+1}.$$ ■

Colin Maclaurin (1698-1746) es más conocido como el defensor del cálculo de Newton cuando éste fue objeto de los ataques implacables del obispo y filósofo irlandés George Berkeley.

Maclaurin no descubrió la serie que lleva su nombre; los matemáticos del siglo ya la conocían desde antes de que él naciera. Sin embargo, concibió un método para resolver un sistema de ecuaciones lineales que se conoce como regla de Cramer, que Cramer no publicó hasta 1750.

Aquí $P_n(x)$ es llamado el ***n*-ésimo polinomio de Taylor** para f alrededor de x_0 y $R_n(x)$ recibe el nombre de **residuo** (o **error de truncamiento**) relacionado con $P_n(x)$. Puesto que el número $\xi(x)$ en el error de truncamiento $R_n(x)$ depende del valor de x donde se evalúa el polinomio $P_n(x)$, es una función de la variable x. Sin embargo, no deberíamos esperar ser capaces de determinar la función $\xi(x)$ de manera explícita. El teorema de Taylor simplemente garantiza que esta función existe y que su valor se encuentra entre x y x_0. De hecho, uno de los problemas comunes en los métodos numéricos es tratar de determinar un límite realista para el valor de $f^{(n+1)}(\xi(x))$ cuando x se encuentra en un intervalo específico.

La serie infinita obtenida al tomar el límite de $P_n(x)$ conforme $n \to \infty$ recibe el nombre de **serie de Taylor** para f alrededor de x_0. En caso de que $x_0 = 0$, entonces al polinomio de Taylor con frecuencia se le llama **polinomio de Maclaurin** y a la serie de Taylor a menudo se le conoce como **serie de Maclaurin**.

El término **error de truncamiento** en el polinomio de Taylor se refiere al error implicado al utilizar una suma truncada, o finita, para aproximar la suma de una serie infinita.

Ejemplo 3 Si $f(x) = \cos x$ y $x_0 = 0$. Determine

a) el segundo polinomio de Taylor para f alrededor de x_0; y

b) el tercer polinomio de Taylor para f alrededor de x_0.

Solución Puesto que $f \in C^\infty(\mathbb{R})$, el teorema de Taylor puede aplicarse a cualquiera $n \geq 0$. Además,

$$f'(x) = -\operatorname{sen} x, \quad f''(x) = -\cos x, \quad f'''(x) = \operatorname{sen} x, \quad \text{y} \quad f^{(4)}(x) = \cos x,$$

Por lo tanto

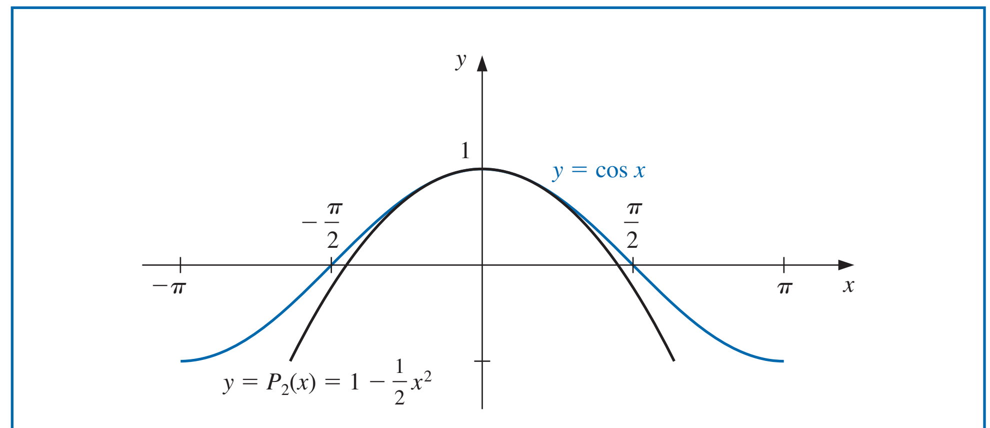
$$f(0) = 1, \quad f'(0) = 0, \quad f''(0) = -1, \quad \text{y} \quad f'''(0) = 0.$$

a) Para $n = 2$ y $x_0 = 0$, obtenemos

$$\begin{aligned}\cos x &= f(0) + f'(0)x + \frac{f''(0)}{2!}x^2 + \frac{f'''(\xi(x))}{3!}x^3 \\ &= 1 - \frac{1}{2}x^2 + \frac{1}{6}x^3 \operatorname{sen} \xi(x),\end{aligned}$$

donde $\xi(x)$ es algún número (por lo general, desconocido) entre 0 y x. (Consulte la figura 1.9.)

Figura 1.9



Cuando $x = 0.01$, esto se convierte en

$$\cos 0.01 = 1 - \frac{1}{2}(0.01)^2 + \frac{1}{6}(0.01)^3 \operatorname{sen} \xi(0.01) = 0.99995 + \frac{10^{-6}}{6} \operatorname{sen} \xi(0.01).$$

Por lo tanto, la aproximación para cos 0.01 provista por el polinomio de Taylor es 0.99995. El error de truncamiento, o término restante, relacionado con esta aproximación es

$$\frac{10^{-6}}{6} \operatorname{sen} \xi(0.01) = 0.1\overline{6} \times 10^{-6} \operatorname{sen} \xi(0.01),$$

donde la barra sobre el 6 en $0.1\overline{6}$ se utiliza para indicar que este dígito se repite indefinidamente. A pesar de que no existe una forma de determinar sen $\xi(0.01)$, sabemos que todos los valores del seno se encuentran en el intervalo $[-1, 1]$, por lo que el error que se presenta si utilizamos la aproximación 0.99995 para el valor de cos 0.01 está limitado por

$$|\cos(0.01) - 0.99995| = 0.1\overline{6} \times 10^{-6} |\operatorname{sen} \xi(0.01)| \leq 0.1\overline{6} \times 10^{-6}.$$

Por lo tanto, la aproximación 0.99995 corresponde por lo menos a los primeros cinco dígitos de cos 0.01 y

$$\begin{aligned} 0.9999483 < 0.99995 - 1.\overline{6} \times 10^{-6} &\leq \cos 0.01 \\ &\leq 0.99995 + 1.\overline{6} \times 10^{-6} < 0.9999517. \end{aligned}$$

El límite del error es mucho más grande que el error real. Esto se debe, en parte, al escaso límite que usamos para $|\operatorname{sen} \xi(x)|$. En el ejercicio 27 se muestra que para todos los valores de x, tenemos $|\operatorname{sen} x| \leq |x|$. Puesto que $0 \leq \xi < 0.01$, podríamos haber usado el hecho de que $|\operatorname{sen} \xi(x)| \leq 0.01$ en la fórmula de error, lo cual produce el límite $0.1\overline{6} \times 10^{-8}$.

b) Puesto que $f'''(0) = 0$, el tercer polinomio de Taylor con el término restante alrededor de $x_0 = 0$ es

$$\cos x = 1 - \frac{1}{2}x^2 + \frac{1}{24}x^4 \cos \tilde{\xi}(x),$$

donde $0 < \tilde{\xi}(x) < 0.01$. El polinomio de aproximación sigue siendo el mismo y la aproximación sigue siendo 0.99995, pero ahora tenemos mayor precisión. Puesto que $|\cos \tilde{\xi}(x)| \leq 1$ para todas las x, obtenemos

$$\left| \frac{1}{24}x^4 \cos \tilde{\xi}(x) \right| \leq \frac{1}{24}(0.01)^4(1) \approx 4.2 \times 10^{-10}.$$

por lo tanto

$$|\cos 0.01 - 0.99995| \leq 4.2 \times 10^{-10},$$

y

$$\begin{aligned} 0.99994999958 = 0.99995 - 4.2 \times 10^{-10} \\ \leq \cos 0.01 \leq 0.99995 + 4.2 \times 10^{-10} = 0.99995000042. \end{aligned}$$ ■

El ejemplo 3 ilustra los dos objetivos del análisis numérico:

i) Encuentre una aproximación a la solución de un problema determinado.

ii) Determine un límite o cota para la precisión de la aproximación.

Los polinomios de Taylor en ambas partes proporcionan la misma respuesta para i), pero el tercero provee una respuesta mucho mejor para ii) que el segundo. También podemos utilizar estos polinomios para obtener aproximaciones de las integrales.

Ilustración Podemos utilizar el tercer polinomio de Taylor y su término restante encontrado en el ejemplo 3 para aproximar $\int_0^{0.1} \cos x\, dx$. Tenemos

$$\begin{aligned}\int_0^{0.1} \cos x\, dx &= \int_0^{0.1} \left(1 - \frac{1}{2}x^2\right) dx + \frac{1}{24}\int_0^{0.1} x^4 \cos \tilde{\xi}(x)\, dx \\ &= \left[x - \frac{1}{6}x^3\right]_0^{0.1} + \frac{1}{24}\int_0^{0.1} x^4 \cos \tilde{\xi}(x)\, dx \\ &= 0.1 - \frac{1}{6}(0.1)^3 + \frac{1}{24}\int_0^{0.1} x^4 \cos \tilde{\xi}(x)\, dx.\end{aligned}$$

Por lo tanto,

$$\int_0^{0.1} \cos x\, dx \approx 0.1 - \frac{1}{6}(0.1)^3 = 0.0998\overline{3}.$$

Un límite o cota para el error en esta aproximación se determina a partir de la integral del término restante de Taylor y el hecho de que $|\cos \tilde{\xi}(x)| \leq 1$ para todas las x:

$$\begin{aligned}\frac{1}{24}\left|\int_0^{0.1} x^4 \cos \tilde{\xi}(x)\, dx\right| &\leq \frac{1}{24}\int_0^{0.1} x^4 |\cos \tilde{\xi}(x)|\, dx \\ &\leq \frac{1}{24}\int_0^{0.1} x^4\, dx = \frac{(0.1)^5}{120} = 8.\overline{3} \times 10^{-8}.\end{aligned}$$

El valor verdadero de esta integral es

$$\int_0^{0.1} \cos x\, dx = \operatorname{sen} x\Big]_0^{0.1} = \operatorname{sen} 0.1 \approx 0.099833416647,$$

por lo que el error real para esta aproximación es 8.3314×10^{-8}, que se encuentra dentro del límite del error. ■

La sección Conjunto de ejercicios 1.1 está disponible en línea. Encuentre la ruta de acceso en las páginas preliminares.

1.2 Errores de redondeo y aritmética computacional

La aritmética realizada con una calculadora o computadora es diferente a la aritmética que se imparte en los cursos de álgebra y cálculo. Podría esperarse que declaraciones como $2+2=4, 4\cdot 8 = 32$, y $(\sqrt{3})^2 = 3$ siempre sean verdaderas; sin embargo con la aritmética *computacional*, esperamos resultados exactos para $2+2=4$ y $4\cdot 8=32$, pero no obtendremos exactamente $(\sqrt{3})^2 = 3$. Para comprender por qué esto es verdadero, debemos explorar el mundo de la aritmética de dígitos finitos.

En nuestro mundo matemático tradicional, permitimos números con una cantidad infinita de dígitos. La aritmética que usamos en este mundo *define* $\sqrt{3}$ como el único número positivo que cuando se multiplica por sí mismo produce el entero 3. No obstante, en el mundo computacional, cada número representable sólo tiene un número fijo y finito de dígitos. Esto significa que, por ejemplo, sólo los números racionales, e incluso no todos ellos, se pueden representar de forma exacta. Ya que $\sqrt{3}$ no es racional, se proporciona una representación aproximada, cuyo cuadrado no será exactamente 3, a pesar de que es probable que esté suficientemente cerca de 3 para ser aceptable en la mayoría de las situaciones. Entonces, en muchos casos, esta aritmética mecánica es satisfactoria y pasa sin importancia o preocupación, pero algunas veces surgen problemas debido a su discrepancia.

El error debido al redondeo debería esperarse siempre que se realizan cálculos con números que no son potencias de 2. Mantener este error bajo control es en extremo importante cuando el número de cálculos es grande.

El error que se produce cuando se utiliza una calculadora o computadora para realizar cálculos con números reales recibe el nombre de **error de redondeo**. Se presenta porque la aritmética realizada en una máquina incluye números con un solo número finito de dígitos, y esto da como resultado cálculos realizados únicamente con representaciones aproximadas de los números reales. En una computadora, sólo un subconjunto relativamente pequeño del sistema de números reales se usa para la representación de todos los números reales. Este subconjunto sólo contiene números racionales, tanto positivos como negativos, y almacena la parte fraccionaria, junto con una parte exponencial.

Números de máquina binarios

En 1985, el Instituto para Ingenieros Eléctricos y Electrónicos (IEEE; Institute for Electrical and Electronic Engineers) publicó un reporte llamado *Binary Floating Point Arithmetic Standard 754–1985 (Estándar para la aritmética binaria de punto flotante)*. En 2008 se publicó una versión actualizada con el nombre de *IEEE 754-2008*; la cual proporciona estándares para números de punto flotante decimales y binarios, formatos para intercambio de datos, algoritmos para redondear operaciones aritméticas y manejo de excepciones. Se especifica cuáles son los formatos para las precisiones individuales, dobles y ampliadas y, en general, todos los fabricantes de microcomputadoras que utilizan hardware de punto flotante siguen estos estándares.

Una representación de 64 bits (dígito binario) se usa para un número real. El primer bit es un indicador de signo, denominado s. A éste le sigue un exponente de 11 bits, c, llamado **característica**, y una fracción binaria de 52 bits, f, llamada **mantisa**. La base para el exponente es 2.

Puesto que los 52 dígitos binarios corresponden con dígitos decimales entre 16 y 17, podemos asumir que un número representado en este sistema tiene por lo menos 16 dígitos decimales de precisión. El exponente de 11 dígitos binarios provee un rango de 0 a $2^{11} - 1 = 2047$. Sin embargo, usar sólo enteros positivos para el exponente no permitiría una representación adecuada de los números con magnitud pequeña. Para garantizar que estos números son igualmente representables, 1023 se resta a la característica, por lo que el rango del exponente en realidad se encuentra entre -1023 y 1024.

Para ahorrar almacenamiento y proporcionar una representación única para cada número de punto flotante, se impone una normalización. Por medio de este sistema obtenemos un número de punto flotante de la forma

$$(-1)^s 2^{c-1023}(1+f).$$

Ilustración Considere el número de máquina

0 10000000011 1011100100010000000000000000000000000000000000000000.

El bit más a la izquierda es $s = 0$, lo cual indica que es un número positivo. Los siguientes 11 bits, 10000000011, proveen la característica y son equivalentes al número decimal

$$c = 1 \cdot 2^{10} + 0 \cdot 2^9 + \cdots + 0 \cdot 2^2 + 1 \cdot 2^1 + 1 \cdot 2^0 = 1024 + 2 + 1 = 1027.$$

La parte exponencial del número es, por lo tanto, $2^{1027-1023} = 2^4$. Los 52 bits finales especifican que la mantisa es

$$f = 1 \cdot \left(\frac{1}{2}\right)^1 + 1 \cdot \left(\frac{1}{2}\right)^3 + 1 \cdot \left(\frac{1}{2}\right)^4 + 1 \cdot \left(\frac{1}{2}\right)^5 + 1 \cdot \left(\frac{1}{2}\right)^8 + 1 \cdot \left(\frac{1}{2}\right)^{12}.$$

Como secuencia, este número de máquina representa precisamente el número decimal

$$(-1)^s 2^{c-1023}(1+f) = (-1)^0 \cdot 2^{1027-1023}\left(1+\left(\frac{1}{2}+\frac{1}{8}+\frac{1}{16}+\frac{1}{32}+\frac{1}{256}+\frac{1}{4096}\right)\right)$$

$$= 27.56640625.$$

Sin embargo, el siguiente número de máquina más pequeño es

$$0\ 10000000011\ 1011100100001111111111111111111111111111111111111111,$$

el siguiente número de máquina más grande es

$$0\ 10000000011\ 1011100100010000000000000000000000000000000000000001.$$

Esto significa que nuestro número de máquina original no sólo representa 27.56640625, sino también la mitad de los números reales que se encuentran entre 27.56640625 y el siguiente número de máquina más pequeño, así como la mitad de los números entre 27.56640625 y el siguiente número de máquina más grande. Para ser preciso, representa cualquier número

$$[27.5664062499999982236431605997495353221893310546875,$$

$$27.5664062500000017763568394002504646778106689453125).$$ ■

El número positivo normalizado más pequeño que se puede representar tiene $s=0$, $c=1$, y $f=0$ y es equivalente

$$2^{-1022} \cdot (1+0) \approx 0.22251 \times 10^{-307},$$

y el más grande tiene $s=0$, $c=2046$ y $f=1-2^{-52}$ y es equivalente a

$$2^{1023} \cdot (2-2^{-52}) \approx 0.17977 \times 10^{309}.$$

Los números que se presentan en los cálculos que tienen una magnitud menor que

$$2^{-1022} \cdot (1+0)$$

resultan en un **subdesbordamiento** y, en general, se configuran en cero. Los números superiores a

$$2^{1023} \cdot (2-2^{-52})$$

resultan en **desbordamiento** y, comúnmente, causan que los cálculos se detengan (a menos que el programa haya sido diseñado para detectar estas presencias). Observe que existen dos representaciones para el número cero: un 0 positivo cuando $s=0$, $c=0$ y $f=0$, y un 0 negativo cuando $s=1$, $c=0$ y $f=0$.

Números de máquina decimales

El uso de dígitos binarios tiende a ocultar las dificultades computacionales que se presentan cuando se usa una colección finita de números de máquina para representar todos los números reales. Para examinar estos problemas, utilizaremos números decimales más familiares en lugar de una representación binaria. En específico, suponemos que los números máquina se representan en formato normalizado de punto flotante *decimal*

$$\pm 0.d_1 d_2 \ldots d_k \times 10^n, \quad 1 \le d_1 \le 9, \quad \text{y} \quad 0 \le d_i \le 9,$$

para cada $i = 2, \ldots, k$. Los números de esta forma reciben el nombre de *números de máquina decimales de dígito k*.

Cualquier número real positivo dentro del rango numérico de la máquina puede ser normalizado a la forma

$$y = 0.d_1 d_2 \ldots d_k d_{k+1} d_{k+2} \ldots \times 10^n.$$

El error que resulta de reemplazar un número con esta forma de punto flotante se llama **error de redondeo**, independientemente de si se usa el método de redondeo o de corte.

La forma de punto flotante de y, que se denota $fl(y)$, se obtiene al terminar la mantisa de y en los dígitos decimales de k. Existen dos maneras comunes para realizar esta terminación. Un método, llamado **de corte**, es simplemente cortar los dígitos $d_{k+1}d_{k+2}\ldots$ Esto produce la forma de punto flotante

$$fl(y) = 0.d_1 d_2 \ldots d_k \times 10^n.$$

El otro método, llamado **redondeo**, suma $5 \times 10^{n-(k+1)}$ a y y entonces corta el resultado para obtener un número con la forma

$$fl(y) = 0.\delta_1 \delta_2 \ldots \delta_k \times 10^n.$$

Para redondear, cuando $d_{k+1} \geq 5$, sumamos 1 a d_k para obtener $fl(y)$; es decir, *redondeamos hacia arriba*. Cuando $d_{k+1} < 5$, simplemente cortamos todo, excepto los primeros dígitos k; es decir, *redondeamos hacia abajo*. Si redondeamos hacia abajo, entonces $\delta_i = d_i$, para cada $i = 1, 2, \ldots, k$. Sin embargo, si redondeamos hacia arriba, los dígitos (e incluso el exponente) pueden cambiar.

Ejemplo 1 Determine los valores a) de corte y b) de redondeo de cinco dígitos del número irracional π.

Solución El número π tiene una expansión decimal infinita de la forma $\pi = 3.14159265\ldots$. Escrito en una forma decimal normalizada, tenemos

$$\pi = 0.314159265\ldots \times 10^1.$$

En general, el error relativo es una mejor medición de precisión que el error absoluto porque considera el tamaño del número que se va a aproximar.

a) El formato de punto flotante de π usando el recorte de cinco dígitos es

$$fl(\pi) = 0.31415 \times 10^1 = 3.1415.$$

b) El sexto dígito de la expansión decimal de π es un 9, por lo que el formato de punto flotante de π con redondeo de cinco dígitos es

$$fl(\pi) = (0.31415 + 0.00001) \times 10^1 = 3.1416.$$ ■

La siguiente definición describe tres métodos para medir errores de aproximación.

Definición 1.15 Suponga que p^* es una aproximación a p. El **error real** es $p - p^*$, el **error absoluto** es $|p - p^*|$, y el **error relativo** es $\dfrac{|p - p^*|}{|p|}$, siempre y cuando $p \neq 0$. ■

Considere los errores real, absoluto y relativo al representar p con p^* en el siguiente ejemplo.

Ejemplo 2 Determine los errores real, absoluto y relativo al aproximar p con p^* cuando

a) $p = 0.3000 \times 10^1$ y $p^* = 0.3100 \times 10^1$;

b) $p = 0.3000 \times 10^{-3}$ y $p^* = 0.3100 \times 10^{-3}$;

c) $p = 0.3000 \times 10^4$ y $p^* = 0.3100 \times 10^4$.

Solución

a) Para $p = 0.3000 \times 10^1$ y $p^* = 0.3100 \times 10^1$, el error real es –0.1, el error absoluto es 0.1 y el error relativo es $0.333\overline{3} \times 10^{-1}$.

b) Para $p = 0.3000 \times 10^{-3}$ y $p^* = 0.3100 \times 10^{-3}$, el error real es -0.1×10^{-4}, el error absoluto es 0.1×10^{-4} y el error relativo es $0.333\overline{3} \times 10^{-1}$.

c) Para $p = 0.3000 \times 10^4$ y $p^* = 0.3100 \times 104$, el error real es -0.1×10^3, el error absoluto es 0.1×10^3 y, de nuevo, el error relativo es $0.333\overline{3} \times 10^{-1}$.

A menudo no podemos encontrar un valor preciso para el error verdadero en una aproximación. Por el contrario, encontramos una cota para el error, lo cual nos proporciona un error del "peor caso".

Este ejemplo muestra que el mismo error relativo, $0.333\overline{3} \times 10^{-1}$, se presenta para errores absolutos ampliamente variables. Como una medida de precisión, el error absoluto puede ser engañoso y el error relativo más significativo debido a que este error considera el tamaño del valor. ■

Un límite de error es un número no negativo mayor que el error absoluto. Algunas veces se obtiene con los métodos de cálculo para encontrar el valor absoluto máximo de una función. Esperamos encontrar el límite superior más pequeño posible para el error a fin de obtener un estimado del error real que es lo más preciso posible.

La siguiente definición usa el error relativo para proporcionar una medida de dígitos significativos de precisión para una aproximación.

Definición 1.16 Se dice que el número p^* se aproxima a p para t **dígitos significativos** (o cifras) si t es el entero no negativo más grande para el que

$$\frac{|p - p^*|}{|p|} \leq 5 \times 10^{-t}.$$ ■

A menudo, el término *dígitos significativos* se usa para describir vagamente el número de dígitos decimales que parecen ser exactos. La definición es más precisa y provee un concepto continuo.

La tabla 1.1 ilustra la naturaleza continua de los dígitos significativos al enumerar, para los diferentes valores de p, el límite superior mínimo de $|p - p^*|$, denominado máx. $|p - p^*|$, cuando p^* concuerda con p en cuatro dígitos significativos.

Tabla 1.1

p	0.1	0.5	100	1000	5000	9990	10000
máx $\|p - p^*\|$	0.00005	0.00025	0.05	0.5	2.5	4.995	5.

Al regresar a la representación de los números de máquina, observamos que la representación de punto flotante $fl(y)$ para el número y tiene el error relativo

$$\left| \frac{y - fl(y)}{y} \right|.$$

Si se usan k dígitos decimales y corte para la representación de máquina de

$$y = 0.d_1 d_2 \ldots d_k d_{k+1} \ldots \times 10^n,$$

entonces

$$\left| \frac{y - fl(y)}{y} \right| = \left| \frac{0.d_1 d_2 \ldots d_k d_{k+1} \ldots \times 10^n - 0.d_1 d_2 \ldots d_k \times 10^n}{0.d_1 d_2 \ldots \times 10^n} \right|$$

$$= \left| \frac{0.d_{k+1} d_{k+2} \ldots \times 10^{n-k}}{0.d_1 d_2 \ldots \times 10^n} \right| = \left| \frac{0.d_{k+1} d_{k+2} \ldots}{0.d_1 d_2 \ldots} \right| \times 10^{-k}.$$

Puesto que $d_1 \neq 0$, el valor mínimo del denominador es 0.1. El numerador se limita en la parte superior mediante 1. Como consecuencia,

$$\left|\frac{y - fl(y)}{y}\right| \leq \frac{1}{0.1} \times 10^{-k} = 10^{-k+1}.$$

De igual forma, un límite para el error relativo al utilizar aritmética de redondeo de dígitos k es $0.5 \times 10^{-k+1}$. (Consulte el ejercicio 28.)

Observe que los límites para el error relativo mediante aritmética de dígitos k son independientes del número que se va a representar. Este resultado se debe a la forma en la que se distribuyen los números de máquina a lo largo de la recta real. Debido al formato exponencial de la característica, el mismo número de los números de máquina decimales se usa para representar cada uno de los intervalos [0.1, 1], [1, 10] y [10, 100]. De hecho, dentro de los límites de la máquina, el número de los números de máquina decimales en $[10^n, 10^{n+1}]$ es constante para todos los enteros n.

Aritmética de dígitos finitos

Además de la representación inexacta de números, la aritmética que se efectúa en una computadora no es exacta. La aritmética implica manipular dígitos binarios mediante diferentes operaciones de cambio, o lógicas. Puesto que la mecánica real de estas operaciones no es pertinente para esta presentación, debemos idear una aproximación propia para aritmética computacional. A pesar de que nuestra aritmética no proporcionará el panorama exacto, es suficiente para explicar los problemas que se presentan. (Para una explicación de las manipulaciones realmente incluidas, se insta al lector a consultar textos de ciencias computacionales con una orientación más técnica, como [Ma], *Computer System Architecture [Arquitectura de sistemas computacionales*.)

Piense que las representaciones de punto flotante $fl(x)$ y $fl(y)$ están dadas para los números reales x y y y que los símbolos $\oplus$, $\ominus$, $\otimes$, y $\oplus\!\!\!\!\div$ representan operaciones de máquina de suma, resta, multiplicación y división, respectivamente. Supondremos una aritmética de dígitos finitos provista por

$$x \oplus y = fl(fl(x) + fl(y)), \quad x \otimes y = fl(fl(x) \times fl(y)),$$
$$x \ominus y = fl(fl(x) - fl(y)), \quad x \oplus\!\!\!\!\div\, y = fl(fl(x) \div fl(y)).$$

Esta aritmética corresponde a realizar aritmética exacta en las representaciones de punto flotante de x y y y, después, convertir el resultado exacto a su representación de punto flotante de dígitos finitos.

Ejemplo 3 Suponga que $x = \frac{5}{7}$ y $y = \frac{1}{3}$. Utilice el corte de cinco dígitos para calcular $x + y$, $x - y$, $x \times y$, y $x \div y$.

Solución Observe que

$$x = \frac{5}{7} = 0.\overline{714285} \quad \text{y} \quad y = \frac{1}{3} = 0.\overline{3}$$

implica que los valores de corte de cinco dígitos de x y y son

$$fl(x) = 0.71428 \times 10^0 \quad \text{y} \quad fl(y) = 0.33333 \times 10^0.$$

Por lo tanto,

$$x \oplus y = fl(fl(x) + fl(y)) = fl\left(0.71428 \times 10^0 + 0.33333 \times 10^0\right)$$
$$= fl\left(1.04761 \times 10^0\right) = 0.10476 \times 10^1.$$

El valor verdadero es $x + y = \frac{5}{7} + \frac{1}{3} = \frac{22}{21}$, por lo que tenemos

$$\text{Error absoluto} = \left| \frac{22}{21} - 0.10476 \times 10^1 \right| = 0.190 \times 10^{-4}$$

y

$$\text{Error relativo} = \left| \frac{0.190 \times 10^{-4}}{22/21} \right| = 0.182 \times 10^{-4}.$$

La tabla 1.2 enumera los valores de éste y otros cálculos. ■

Tabla 1.2

Operación	Resultado	Valor real	Error absoluto	Error relativo
$x \oplus y$	0.10476×10^1	22/21	0.190×10^{-4}	0.182×10^{-4}
$x \ominus y$	0.38095×10^0	8/21	0.238×10^{-5}	0.625×10^{-5}
$x \otimes y$	0.23809×10^0	5/21	0.524×10^{-5}	0.220×10^{-4}
$x \oslash y$	0.21428×10^1	15/7	0.571×10^{-4}	0.267×10^{-4}

El error relativo máximo para las operaciones en el ejemplo 3 es 0.267×10^{-4}, por lo que la aritmética produce resultados satisfactorios de cinco dígitos. Éste no es el caso en el siguiente ejemplo.

Ejemplo 4 Suponga que además de $x = \frac{5}{7}$ y $y = \frac{1}{3}$ tenemos

$$u = 0.714251, \quad v = 98765.9, \quad \text{y} \quad w = 0.111111 \times 10^{-4},$$

de tal forma que

$$fl(u) = 0.71425 \times 10^0, \quad fl(v) = 0.98765 \times 10^5, \quad \text{y} \quad fl(w) = 0.11111 \times 10^{-4}.$$

Determine los valores de corte de cinco dígitos de $x \ominus u$, $(x \ominus u) \oslash w$, $(x \ominus u) \otimes v$, y $u \oplus v$.

Solución Estos números fueron seleccionados para ilustrar algunos problemas que pueden surgir con la aritmética de dígitos finitos. Puesto que x y u son casi iguales, su diferencia es pequeña. El error absoluto para $x \ominus u$ es

$$\begin{aligned}
|(x - u) - (x \ominus u)| &= |(x - u) - (fl(fl(x) - fl(u)))| \\
&= \left| \left(\frac{5}{7} - 0.714251 \right) - \left(fl\left(0.71428 \times 10^0 - 0.71425 \times 10^0\right)\right) \right| \\
&= \left|0.347143 \times 10^{-4} - fl\left(0.00003 \times 10^0\right)\right| = 0.47143 \times 10^{-5}.
\end{aligned}$$

Esta aproximación tiene un error absoluto, pero un error relativo grande

$$\left| \frac{0.47143 \times 10^{-5}}{0.347143 \times 10^{-4}} \right| \leq 0.136.$$

La división subsiguiente entre el número pequeño w o la multiplicación por el número grande v magnifica el error absoluto sin modificar el error relativo. La suma de los números grande y pequeño u y v produce un error absoluto grande, pero no un error relativo grande. Estos cálculos se muestran en la tabla 1.3. ■

Tabla 1.3

Operación	Resultado	Valor real	Error absoluto	Error relativo
$x \ominus u$	0.30000×10^{-4}	0.34714×10^{-4}	0.471×10^{-5}	0.136
$(x \ominus u) \oplus w$	0.27000×10^{1}	0.31242×10^{1}	0.424	0.136
$(x \ominus u) \otimes v$	0.29629×10^{1}	0.34285×10^{1}	0.465	0.136
$u \oplus v$	0.98765×10^{5}	0.98766×10^{5}	0.161×10^{1}	0.163×10^{-4}

Uno de los cálculos más comunes que producen errores implica la cancelación de dígitos significativos debido a la resta de números casi iguales. Suponga que dos números casi iguales x y y, con $x > y$, tienen las representaciones de dígitos k

$$fl(x) = 0.d_1d_2 \ldots d_p\alpha_{p+1}\alpha_{p+2} \ldots \alpha_k \times 10^n$$

y

$$fl(y) = 0.d_1d_2 \ldots d_p\beta_{p+1}\beta_{p+2} \ldots \beta_k \times 10^n.$$

El formato de punto flotante de $x - y$ es

$$fl(fl(x) - fl(y)) = 0.\sigma_{p+1}\sigma_{p+2} \ldots \sigma_k \times 10^{n-p},$$

donde

$$0.\sigma_{p+1}\sigma_{p+2} \ldots \sigma_k = 0.\alpha_{p+1}\alpha_{p+2} \ldots \alpha_k - 0.\beta_{p+1}\beta_{p+2} \ldots \beta_k.$$

El número de punto flotante que se usa para representar $x - y$ tiene por lo menos $k - p$ dígitos significativos. Sin embargo, en muchos dispositivos de cálculo a $x - y$ se le asignarán k dígitos, con la última p igual a cero o asignada de manera aleatoria. Cualquier otro cálculo relacionado con $x - y$ conserva el problema de tener solamente $k - p$ dígitos significativos, puesto que una cadena de cálculos no es más precisa que su parte más débil.

Si una representación o un cálculo de dígitos finitos presenta un error, otra ampliación del error ocurre al dividir entre un número de menor magnitud (o, de manera equivalente, al multiplicar por un número de mayor magnitud). Suponga, por ejemplo, que el número z tiene una aproximación de dígitos finitos $z + \delta$, en donde el error δ se introduce por representación o por cálculo previo. Ahora divida entre $\varepsilon = 10^{-n}$, en donde $n > 0$. Entonces

$$\frac{z}{\varepsilon} \approx fl\left(\frac{fl(z)}{fl(\varepsilon)}\right) = (z + \delta) \times 10^n.$$

El error absoluto en esta aproximación, $|\delta| \times 10^n$, es el error absoluto original, $|\delta|$, multiplicado por el factor 10^n.

Ejemplo 5 Si $p = 0.54617$ y $q = 0.54601$. Use aritmética de cuatro dígitos para aproximar $p - q$ y determine los errores absoluto y relativo mediante **a)** redondeo y **b)** corte.

Solución El valor exacto de $r = p - q$ es $r = 0.00016$.

a) Suponga que se realiza la resta con aritmética de redondeo de cuatro dígitos. Al redondear p y q a cuatro dígitos obtenemos $p^* = 0.5462$ y $q^* = 0.5460$, respectivamente, y $r^* = p^* - q^* = 0.0002$ es la aproximación de cuatro dígitos para r. Puesto que

$$\frac{|r - r^*|}{|r|} = \frac{|0.00016 - 0.0002|}{|0.00016|} = 0.25,$$

el resultado sólo tiene un dígito significativo, mientras p^* y q^* sean precisos para cuatro y cinco dígitos significativos, respectivamente.

b) Si se usa el corte para obtener los cuatro dígitos, la aproximación de cuatro dígitos para p, q, y r son $p^* = 0.5461$, $q^* = 0.5460$, y $r^* = p^* - q^* = 0.0001$. Esto nos da

$$\frac{|r - r^*|}{|r|} = \frac{|0.00016 - 0.0001|}{|0.00016|} = 0.375,$$

lo que también resulta en un solo dígito significativo de precisión. ■

A menudo, la pérdida de precisión debido al error de redondeo se puede evitar al reformular los cálculos, como se ilustra en el siguiente ejemplo.

Ilustración La fórmula cuadrática establece que las raíces de $ax^2 + bx + c = 0$, cuando $a \neq 0$, son

$$x_1 = \frac{-b + \sqrt{b^2 - 4ac}}{2a} \quad \text{y} \quad x_2 = \frac{-b - \sqrt{b^2 - 4ac}}{2a}. \tag{1.1}$$

Considere esta fórmula aplicada a la ecuación $x^2 + 62.10x + 1 = 0$, cuyas raíces son aproximadamente

$$x_1 = -0.01610723 \quad \text{y} \quad x_2 = -62.08390.$$

Las raíces x_1 y x_2 de una ecuación cuadrática general están relacionadas con los coeficientes por el hecho de que

$$x_1 + x_2 = -\frac{b}{a} \quad \text{y} \quad x_1 x_2 = \frac{c}{a}.$$

Éste es un caso especial de las fórmulas de Vièta para los coeficientes de los polinomios

Usaremos otra vez la aritmética de redondeo de cuatro dígitos en los cálculos para determinar la raíz. En esta ecuación, b^2 es mucho más grande que $4ac$, por lo que el numerador en el cálculo para x_1 implica la *resta* de números casi iguales. Ya que

$$\sqrt{b^2 - 4ac} = \sqrt{(62.10)^2 - (4.000)(1.000)(1.000)}$$
$$= \sqrt{3856. - 4.000} = \sqrt{3852.} = 62.06,$$

tenemos

$$fl(x_1) = \frac{-62.10 + 62.06}{2.000} = \frac{-0.04000}{2.000} = -0.02000,$$

una aproximación deficiente a $x_1 = -0.01611$, con un error relativo grande

$$\frac{|-0.01611 + 0.02000|}{|-0.01611|} \approx 2.4 \times 10^{-1}.$$

Por otro lado, el cálculo para x_2 implica la suma de los números casi iguales $-b$ y $-\sqrt{b^2 - 4ac}$. Esto no presenta problemas debido a que

$$fl(x_2) = \frac{-62.10 - 62.06}{2.000} = \frac{-124.2}{2.000} = -62.10$$

tiene un error relativo pequeño

$$\frac{|-62.08 + 62.10|}{|-62.08|} \approx 3.2 \times 10^{-4}.$$

Para obtener una aproximación por redondeo de cuatro dígitos para x_1, modificamos el formato de la fórmula cuadrática al *racionalizar el numerador*:

$$x_1 = \frac{-b + \sqrt{b^2 - 4ac}}{2a} \left(\frac{-b - \sqrt{b^2 - 4ac}}{-b - \sqrt{b^2 - 4ac}} \right) = \frac{b^2 - (b^2 - 4ac)}{2a(-b - \sqrt{b^2 - 4ac})},$$

la cual se simplifica en una fórmula cuadrática alterna:

$$x_1 = \frac{-2c}{b + \sqrt{b^2 - 4ac}}. \tag{1.2}$$

Por medio de la ecuación (1.2) obtenemos

$$fl(x_1) = \frac{-2.000}{62.10 + 62.06} = \frac{-2.000}{124.2} = -0.01610,$$

que tiene el error relativo pequeño 6.2×10^{-4}.

La técnica de racionalización también puede aplicarse para proporcionar la siguiente fórmula cuadrática alterna para x_2:

$$x_2 = \frac{-2c}{b - \sqrt{b^2 - 4ac}}. \tag{1.3}$$

Éste es el formato que se usa si b es un número negativo. En la ilustración, sin embargo, el uso erróneo de esta fórmula para x_2 no sólo resultaría en la resta de números casi iguales, sino también en la división entre el resultado pequeño de esta resta. La falta de precisión que produce esta combinación,

$$fl(x_2) = \frac{-2c}{b - \sqrt{b^2 - 4ac}} = \frac{-2.000}{62.10 - 62.06} = \frac{-2.000}{0.04000} = -50.00,$$

tiene el error relativo grande 1.9×10^{-1}. ■

- La lección: ¡Piense antes de calcular!

Aritmética anidada

La pérdida de precisión debido a un error de redondeo también se puede reducir al reacomodar los cálculos, como se muestra en el siguiente ejemplo.

Ejemplo 6 Evalúe $f(x) = x^3 - 6.1x^2 + 3.2x + 1.5$ en $x = 4.71$ con aritmética de tres dígitos.

Solución La tabla 1.4 provee los resultados intermedios de los cálculos.

Tabla 1.4

	x	x^2	x^3	$6.1x^2$	$3.2x$
Exacto	4.71	22.1841	104.487111	135.32301	15.072
Tres dígitos (corte)	4.71	22.1	104.	134.	15.0
Tres dígitos (redondeo)	4.71	22.2	105.	135.	15.1

Para ilustrar los cálculos, observemos los que participan para encontrar x^3 usando la aritmética de redondeo de tres dígitos. Primero encontramos

$$x^2 = 4.71^2 = 22.1841 \quad \text{que se redondea a } 22.2.$$

A continuación, usamos este valor de x^2 para encontrar

$$x^3 = x^2 \cdot x = 22.2 \cdot 4.71 = 104.562 \quad \text{que se redondea a } 105.$$

Además,

$$6.1x^2 = 6.1(22.2) = 135.42 \quad \text{que se redondea a } 135,$$

y

$$3.2x = 3.2(4.71) = 15.072 \quad \text{que se redondea a 15.1.}$$

El resultado exacto de la evaluación es

$$\text{Exacto: } f(4.71) = 104.487111 - 135.32301 + 15.072 + 1.5 = -14.263899.$$

Con la aritmética de dígitos finitos, la forma en la que sumamos los resultados puede afectar el resultado final. Suponga que lo hacemos de izquierda a derecha. Entonces, para la aritmética de corte tenemos

$$\text{Tres dígitos (corte): } f(4.71) = ((104. - 134.) + 15.0) + 1.5 = -13.5,$$

y para la aritmética de redondeo tenemos

$$\text{Tres dígitos (redondeo): } f(4.71) = ((105. - 135.) + 15.1) + 1.5 = -13.4.$$

(Usted debe verificar de manera cuidadosa estos resultados para asegurarse de que su noción de aritmética de dígitos finitos es correcta.) Observe que los valores de corte de tres dígitos sólo retienen los tres dígitos principales, sin incluir redondeo, y difieren significativamente de los valores de redondeo de tres dígitos.

Los errores relativos para los métodos de tres dígitos son

$$\text{Corte: } \left|\frac{-14.263899 + 13.5}{-14.263899}\right| \approx 0.05, \quad \text{y redondeo: } \left|\frac{-14.263899 + 13.4}{-14.263899}\right| \approx 0.06.$$

■

Ilustración

Recuerde que el corte (o redondeo) se realiza después del cálculo.

Como enfoque alternativo, el polinomio $f(x)$ en el ejemplo 6 se puede reescribir de forma **anidada** como

$$f(x) = x^3 - 6.1x^2 + 3.2x + 1.5 = ((x - 6.1)x + 3.2)x + 1.5.$$

Ahora, la aritmética de corte de tres dígitos produce

$$\begin{aligned} f(4.71) =& ((4.71 - 6.1)4.71 + 3.2)4.71 + 1.5 = ((-1.39)(4.71) + 3.2)4.71 + 1.5 \\ =& (-6.54 + 3.2)4.71 + 1.5 = (-3.34)4.71 + 1.5 = -15.7 + 1.5 = -14.2. \end{aligned}$$

De manera similar, ahora obtenemos una respuesta de redondeo de tres dígitos de -14.3. Los nuevos errores relativos son

$$\text{Tres dígitos (corte): } \left|\frac{-14.263899 + 14.2}{-14.263899}\right| \approx 0.0045;$$

$$\text{Tres dígitos (redondeo): } \left|\frac{-14.263899 + 14.3}{-14.263899}\right| \approx 0.0025.$$

El anidado ha reducido el error relativo para la aproximación de corte a menos de 10% del valor obtenido al inicio. Para la aproximación de redondeo, la mejora ha sido todavía más drástica; el error, en este caso, se ha reducido más de 95 por ciento. ■

Los polinomios *siempre* deberían expresarse en forma anidada antes de realizar una evaluación porque esta forma minimiza el número de cálculos aritméticos. La disminución del error en la ilustración se debe a la reducción de los cálculos de cuatro multiplicaciones y tres sumas a dos multiplicaciones y tres sumas. Una forma de disminuir el error de redondeo es reducir el número de cálculos.

La sección Conjunto de ejercicios 1.2 está disponible en línea. Encuentre la ruta de acceso en las páginas preliminares.

1.3 Algoritmos y convergencia

A lo largo del texto examinaremos procedimientos de aproximación, llamados *algoritmos*, los cuales incluyen secuencias de cálculos. Un **algoritmo** es un procedimiento que describe, de manera inequívoca, una secuencia finita de pasos que se desarrollarán en un orden específico. El objeto del algoritmo es implementar un procedimiento para resolver un problema o aproximar una solución para el problema.

El uso de algoritmos es tan antiguo como las matemáticas formales pero el nombre se deriva del matemático árabe Muhammad ibn-Msâ al-Khwarârizmî (c. 780–850). La traducción latina de sus trabajos comenzó con las palabras "Dixit Algorismi", que significan "al-Khwarârizmî dice".

Nosotros usamos un **pseudocódigo** para describir los algoritmos. Este pseudocódigo describe la forma de la entrada que se proporcionará y la forma de la salida deseada. No todos los procedimientos proporcionan una salida satisfactoria para una entrada seleccionada de manera arbitraria. Como consecuencia, se incluye una técnica de detención independiente de la técnica numérica en cada algoritmo para evitar ciclos infinitos.

En los algoritmos se usan dos símbolos de puntuación:

- Un punto (.) indica el final de un paso.
- Punto y coma (;) separan las tareas dentro de un paso.

La sangría se usa para indicar que los grupos de declaraciones se tratarán como una sola entidad.

Las técnicas de ciclado en los algoritmos también son controladas por contador, como

Para $i = 1, 2, \ldots, n$
tome $x_i = a + i \cdot h$

O controladas por condición, como

Mientras $i < N$ realice los pasos 3–6.

Para permitir una ejecución condicional, usamos las construcciones estándar

Si… entonces o Si… entonces
si no

Los pasos en los algoritmos siguen las reglas de la construcción del programa estructurado. Se han ordenado de tal forma que no debería ser difícil traducir el pseudocódigo en cualquier lenguaje de programación adecuado para aplicaciones científicas.

Los algoritmos están mezclados libremente con comentarios. Éstos se escriben en itálicas y se encuentran entre paréntesis para distinguirlos de las declaraciones algorítmicas.

NOTA: Cuando es difícil determinar el final de ciertos pasos anidados utilizamos un comentario como (fin del paso 14) a la derecha o debajo de la declaración de finalización. Consulte, por ejemplo, el comentario en el paso 5, en el ejemplo 1.

Ilustración El siguiente algoritmo calcula $x_1 + x_2 + \cdots + x_N = \sum_{i=1}^{N} x_i$, dado N y los números $x_1, x_2, \ldots, x_N$

ENTRADA $N, x_1, x_2, \ldots, x_n$.

SALIDA $SUM = \sum_{i=1}^{N} x_i$.

Paso 1 Tome $SUM = 0$. (*Inicialice el acumulador*.)

Paso 2 Para $i = 1, 2, \ldots, N$ hacer
Tome $SUM = SUM + x_i$. (*Añadir el siguiente término*.)

Paso 3 SALIDA (SUM);
PARE. ■

Ejemplo 1 El n-ésimo polinomio de Taylor para $f(x) = \ln x$ ampliado alrededor de $x_0 = 1$ es

$$P_N(x) = \sum_{i=1}^{N} \frac{(-1)^{i+1}}{i}(x-1)^i,$$

y el valor de ln 1.5 para los ocho lugares decimales es 0.40546511. Construya un algoritmo para determinar el valor mínimo de N requerido para

$$|\ln 1.5 - P_N(1.5)| < 10^{-5}$$

sin utilizar el término restante del polinomio de Taylor.

Solución A partir del cálculo sabemos que si $\sum_{n=1}^{\infty} a_n$ es una serie alterna con límite A cuyos términos disminuyen en magnitud, entonces A y la n-ésima suma parcial $A_N = \sum_{n=1}^{N} a_n$ difieren por menos en la magnitud del término $(N+1)$; es decir,

$$|A - A_N| \leq |a_{N+1}|.$$

El siguiente algoritmo utiliza este hecho.

ENTRADA valor x, tolerancia TOL, número máximo de iteraciones M.
SALIDA grado N del polinomio o un mensaje de falla.
Paso 1 Sea $N = 1$;
$y = x - 1$;
$SUM = 0$;
$POWER = y$;
$TERM = y$;
$SIGN = -1$. (*Se utiliza para implementar la alternancia de los signos*.)

Paso 2 Mientras $N \leq M$ haga los pasos 3–5.

Paso 3 Determine $SIGN = -SIGN$; (*Alterne los signos*.)
$SUM = SUM + SIGN \cdot TERM$; (*Acumule los términos*.)
$POWER = POWER \cdot y$;
$TERM = POWER/(N+1)$. (*Calcule el siguiente término*.)

Paso 4 Si $|TERM| < TOL$ entonces (*Prueba para la precisión*.)
SALIDA (N);
PARE. (*El procedimiento fue exitoso*.)

Paso 5 Determinar $N = N + 1$. (*Preparar la siguiente iteración*. (*Fin del paso 2*))

Paso 6 SALIDA ('El método falló'); (*El procedimiento no fue exitoso*.)
PARE.

La entrada para nuestro problema es $x = 1.5$, $TOL = 10^{-5}$ y tal vez $M = 15$. Esta elección de M provee un límite o una cota superior para el número de cálculos que queremos realizar, al reconocer que el algoritmo probablemente va a fallar si se excede este límite. La salida es un valor para N o el mensaje de fracaso que depende de la precisión del dispositivo computacional. ■

Algoritmos de caracterización

Consideraremos diversos problemas de aproximación a lo largo del texto y en cada caso necesitamos determinar métodos de aproximación que producen resultados precisos fiables para una amplia clase de problemas. Debido a las diferentes formas de derivar los métodos de aproximación, requerimos una variedad de condiciones para clasificar su precisión. No todas estas condiciones son apropiadas para cualquier problema en particular.

La palabra *estable* tiene la misma raíz que las palabras *posición* y *estándar.* En matemáticas, el término *estable* aplicado a un problema indica que un pequeño cambio en los datos o las condiciones iniciales no resultan en un cambio drástico en la solución del problema.

Un criterio que impondremos en un algoritmo, siempre que sea posible, es que los pequeños cambios en los datos iniciales producen, de forma proporcional, pequeños cambios en los resultados finales. Un algoritmo que satisface esta propiedad recibe el nombre de **estable**; de lo contrario, es **inestable**. Algunos algoritmos son estables sólo para ciertas elecciones de datos iniciales y reciben el nombre de **estables condicionalmente**. Clasificaremos las propiedades de estabilidad de los algoritmos siempre que sea posible.

Para considerar más el tema del crecimiento del error de redondeo y su conexión con la estabilidad del algoritmo, suponga que se presenta un error con una magnitud $E_0 > 0$ en alguna etapa en los cálculos y que la magnitud del error después de n operaciones subsiguientes se denota con E_n. En la práctica, los dos casos que surgen con mayor frecuencia se definen a continuación.

Definición 1.17 Suponga que $E_0 > 0$ denota un error que se presenta en alguna etapa en los cálculos y E_n representa la magnitud del error después de n operaciones subsiguientes.

- Si $E_n \approx C^n E_0$, donde C es una constante independiente de n, entonces se dice que el crecimiento del error es **lineal**.
- Si $E_n \approx C^n E_0$, para algunas $C > 1$, entonces el crecimiento del error recibe el nombre de **exponencial**. ■

Normalmente, el crecimiento lineal del error es inevitable, y cuando C y E_0 son pequeñas, en general, los resultados son aceptables. El crecimiento exponencial del error debería evitarse porque el término C^n se vuelve grande incluso para los valores relativamente pequeños de n. Esto conduce a imprecisiones inaceptables, independientemente del tamaño de E_0. Como consecuencia, mientras un algoritmo que presenta crecimiento lineal del error es estable, un algoritmo que presenta crecimiento exponencial del error es inestable. (Consulte la figura 1.10.)

Figura 1.10

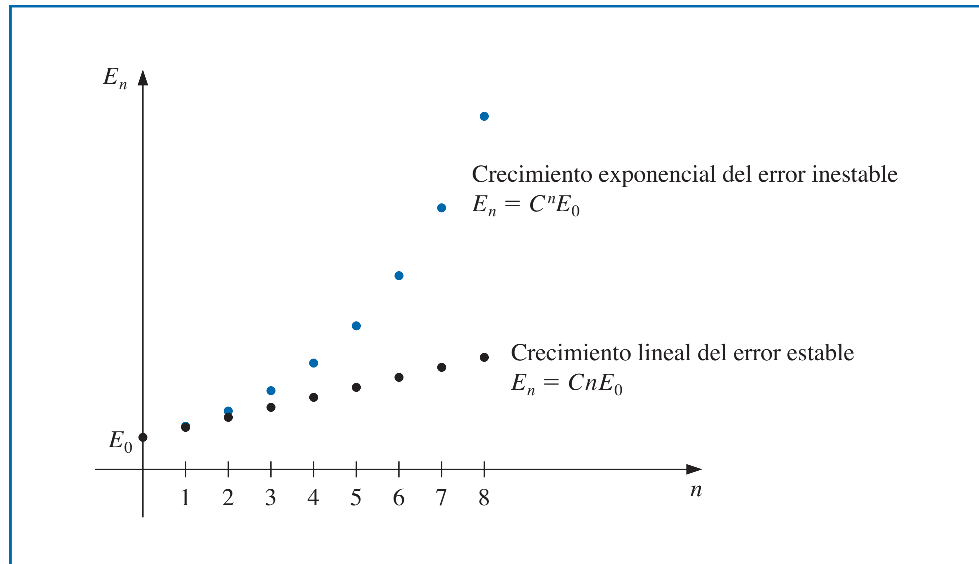


Ilustración Para cualquier constante c_1 y c_2,

$$p_n = c_1 \left(\frac{1}{3}\right)^n + c_2 3^n, \tag{1.4}$$

es una solución a la ecuación recursiva

$$p_n = \frac{10}{3} p_{n-1} - p_{n-2}, \quad \text{para } n = 2, 3, \ldots .$$

Se puede observar que

$$\begin{aligned}
\frac{10}{3} p_{n-1} - p_{n-2} &= \frac{10}{3}\left[c_1\left(\frac{1}{3}\right)^{n-1} + c_2 3^{n-1}\right] - \left[c_1\left(\frac{1}{3}\right)^{n-2} + c_2 3^{n-2}\right] \\
&= c_1\left(\frac{1}{3}\right)^{n-2}\left[\frac{10}{3}\cdot\frac{1}{3} - 1\right] + c_2 3^{n-2}\left[\frac{10}{3}\cdot 3 - 1\right] \\
&= c_1\left(\frac{1}{3}\right)^{n-2}\left(\frac{1}{9}\right) + c_2 3^{n-2}(9) = c_1\left(\frac{1}{3}\right)^{n} + c_2 3^n = p_n.
\end{aligned}$$

Suponga que tenemos $p_0 = 1$ y $p_1 = \frac{1}{3}$. Usando estos valores y la ecuación (1.4) podemos determinar valores únicos para las constantes $c_1 = 1$ y $c_2 = 0$. Por lo tanto, $p_n = \left(\frac{1}{3}\right)^n$ para todas las n.

Si se utiliza aritmética de redondeo de cinco dígitos para calcular los términos de la sucesión determinada por esta ecuación, entonces $\hat{p}_0 = 1.0000$ y $\hat{p}_1 = 0.33333$, lo cual requiere la modificación de las constantes para $\hat{c}_1 = 1.0000$ y $\hat{c}_2 = -0.12500 \times 10^{-5}$. Así, la sucesión generada $\{\hat{p}_n\}_{n=0}^{\infty}$ está dada por

$$\hat{p}_n = 1.0000\left(\frac{1}{3}\right)^n - 0.12500 \times 10^{-5}(3)^n,$$

que tiene un error de redondeo,

$$p_n - \hat{p}_n = 0.12500 \times 10^{-5}(3^n).$$

Este procedimiento es inestable ya que el error aumenta *exponencialmente* con n, lo cual se refleja en las imprecisiones extremas después de los primeros términos, como se muestra en la tabla 1.5.

Tabla 1.5

n	$\hat{p}_n$ calculada	p_n corregida	Error relativo
0	0.10000×10^{1}	0.10000×10^{1}	
1	0.33333×10^{0}	0.33333×10^{0}	
2	0.11110×10^{0}	0.11111×10^{0}	9×10^{-5}
3	0.37000×10^{-1}	0.37037×10^{-1}	1×10^{-3}
4	0.12230×10^{-1}	0.12346×10^{-1}	9×10^{-3}
5	0.37660×10^{-2}	0.41152×10^{-2}	8×10^{-2}
6	0.32300×10^{-3}	0.13717×10^{-2}	8×10^{-1}
7	-0.26893×10^{-2}	0.45725×10^{-3}	7×10^{0}
8	-0.92872×10^{-2}	0.15242×10^{-3}	6×10^{1}

Ahora considere esta ecuación recursiva:

$$p_n = 2p_{n-1} - p_{n-2}, \quad \text{para } n = 2, 3, \ldots .$$

Tiene la solución $p_n = c_1 + c_2 n$ para cualquier constante c_1 y c_2 porque

$$\begin{aligned}
2p_{n-1} - p_{n-2} &= 2(c_1 + c_2(n-1)) - (c_1 + c_2(n-2)) \\
&= c_1(2-1) + c_2(2n - 2 - n + 2) = c_1 + c_2 n = p_n.
\end{aligned}$$

Si tenemos $p_0 = 1$ y $p_1 = \frac{1}{3}$, entonces las constantes en esta ecuación se determinan exclusivamente como $c_1 = 1$ y $c_2 = -\frac{2}{3}$. Esto implica que $p_n = 1 - \frac{2}{3}n$.

Si se utiliza aritmética de redondeo de cinco dígitos para calcular los términos de la sucesión provista por esta ecuación, entonces $\hat{p}_0 = 1.0000$ y $\hat{p}_1 = 0.33333$. Como consecuencia, las constantes de redondeo de cinco dígitos son $\hat{c}_1 = 1.0000$ y $\hat{c}_2 = -0.66667$. Por lo tanto,

$$\hat{p}_n = 1.0000 - 0.66667n,$$

que tiene un error de redondeo

$$p_n - \hat{p}_n = \left(0.66667 - \frac{2}{3}\right) n.$$

Este procedimiento es estable porque el error aumenta *linealmente* con n, lo cual se refleja en las aproximaciones que se muestran en la tabla 1.6. ■

Tabla 1.6

n	$\hat{p}_n$ calculada	p_n corregida	Error relativo
0	0.10000×10^1	0.10000×10^1	
1	0.33333×10^0	0.33333×10^0	
2	-0.33330×10^0	-0.33333×10^0	9×10^{-5}
3	-0.10000×10^1	-0.10000×10^1	0
4	-0.16667×10^1	-0.16667×10^1	0
5	-0.23334×10^1	-0.23333×10^1	4×10^{-5}
6	-0.30000×10^1	-0.30000×10^1	0
7	-0.36667×10^1	-0.36667×10^1	0
8	-0.43334×10^1	-0.43333×10^1	2×10^{-5}

Los efectos del error de redondeo se pueden reducir con la aritmética de dígitos de orden superior, como la opción de precisión doble o múltiple disponible en muchas computadoras. Las desventajas de utilizar la aritmética de precisión doble son que requiere más tiempo de cálculo y el crecimiento del error de redondeo no se elimina por completo.

Un enfoque para calcular el error de redondeo es usar la aritmética de intervalo (es decir, retener los valores más grande y más pequeño posibles), de esta forma, al final, obtenemos un intervalo que contiene el valor verdadero. Por desgracia, podría ser necesario un intervalo pequeño para la implementación razonable.

Tasas de convergencia

Puesto que con frecuencia se utilizan técnicas iterativas relacionadas con sucesiones, esta sección concluye con un análisis breve sobre la terminología que se usa para describir la rapidez con que ocurre la convergencia. En general, nos gustaría que la técnica converja tan rápido como sea posible. La siguiente definición se usa para comparar las tasas de convergencia de las sucesiones.

Definición 1.18 Suponga que $\{\beta_n\}_{n=1}^{\infty}$ es una sucesión conocida que converge a cero y $\{\alpha_n\}_{n=1}^{\infty}$ converge a un número α. Si existe una constante positiva K con

$$|\alpha_n - \alpha| \leq K|\beta_n|, \qquad \text{para una } n \text{ grande,}$$

entonces decimos que $\{\alpha_n\}_{n=1}^{\infty}$ converge a α con **una rapidez, u orden de convergencia** $O(\beta_n)$. (Esta expresión se lee "O de β_n".) Se indica al escribir $\alpha_n = \alpha + O(\beta_n)$. ■

A pesar de que la definición 1.18 permite comparar $\{\alpha_n\}_{n=1}^{\infty}$ con una sucesión arbitraria $\{\beta_n\}_{n=1}^{\infty}$, en casi todas las situaciones usamos

$$\beta_n = \frac{1}{n^p},$$

para algún número $p > 0$. En general, nos interesa el valor más grande de p con $\alpha_n = \alpha + O(1/n^p)$.

Ejemplo 2 Suponga que, para $n \geq 1$,

$$\alpha_n = \frac{n+1}{n^2} \quad \text{y} \quad \hat{\alpha}_n = \frac{n+3}{n^3}.$$

aunque $\lim_{n\to\infty} \alpha_n = 0$ y $\lim_{n\to\infty} \hat{\alpha}_n = 0$, la sucesión $\{\hat{\alpha}_n\}$ converge a este límite mucho más rápido que la sucesión $\{\alpha_n\}$. Al usar la aritmética de redondeo de cinco dígitos, tenemos los valores que se muestran en la tabla 1.7. Determine la rapidez de convergencia para estas dos sucesiones.

Tabla 1.7

n	1	2	3	4	5	6	7
α_n	2.00000	0.75000	0.44444	0.31250	0.24000	0.19444	0.16327
$\hat{\alpha}_n$	4.00000	0.62500	0.22222	0.10938	0.064000	0.041667	0.029155

Existen muchas otras formas de describir el crecimiento de las sucesiones y las funciones, algunas requieren límites tanto por encima como por debajo de la sucesión o función que se considera. Cualquier buen libro que analiza algoritmos, por ejemplo, [CLRS], incluiría esta información.

Solución Defina las sucesiones $\beta_n = 1/n$ y $\hat{\beta}_n = 1/n^2$. Entonces

$$|\alpha_n - 0| = \frac{n+1}{n^2} \leq \frac{n+n}{n^2} = 2 \cdot \frac{1}{n} = 2\beta_n$$

y

$$|\hat{\alpha}_n - 0| = \frac{n+3}{n^3} \leq \frac{n+3n}{n^3} = 4 \cdot \frac{1}{n^2} = 4\hat{\beta}_n.$$

Por lo tanto, la rapidez de convergencia de $\{\alpha_n\}$ cero es similar a la convergencia de $\{1/n\}$ a cero, mientras $\{\hat{\alpha}_n\}$ converge a cero con una rapidez similar para la sucesión que converge más rápido $\{1/n^2\}$. Expresamos esto al escribir

$$\alpha_n = 0 + O\left(\frac{1}{n}\right) \quad \text{y} \quad \hat{\alpha}_n = 0 + O\left(\frac{1}{n^2}\right).$$

■

También usamos la notación O (O *grande*) para describir la rapidez con la que convergen las funciones.

Definición 1.19 Suponga que $\lim_{h\to 0} G(h) = 0$ y $\lim_{h\to 0} F(h) = L$. Si existe una constante positiva K con

$$|F(h) - L| \leq K|G(h)|, \qquad \text{para } h \text{ suficientemente pequeña},$$

entonces escribimos $F(h) = L + O(G(h))$. ■

En general, las funciones que utilizamos para comparar tienen la forma de $G(h) = h^p$, donde $p > 0$. Nos interesa el valor más grande de p, para el que $F(h) = L + O(h^p)$.

Ejemplo 3 Use el tercer polinomio de Taylor alrededor de $h = 0$ para mostrar que $\cos h + \frac{1}{2}h^2 = 1 + O(h^4)$.

Solución En el ejemplo 3b) de la sección 1.1, vimos que este polinomio es

$$\cos h = 1 - \frac{1}{2}h^2 + \frac{1}{24}h^4 \cos \tilde{\xi}(h),$$

para algún número $\tilde{\xi}(h)$ entre cero y h. Esto implica que

$$\cos h + \frac{1}{2}h^2 = 1 + \frac{1}{24}h^4 \cos \tilde{\xi}(h).$$

Por lo tanto,

$$\left| \left(\cos h + \frac{1}{2}h^2 \right) - 1 \right| = \left| \frac{1}{24} \cos \tilde{\xi}(h) \right| h^4 \leq \frac{1}{24}h^4,$$

de modo que $h \to 0$, $\cos h + \frac{1}{2}h^2$ converge a este límite, 1, tan rápido como h^4 converge a 0. Es decir,

$$\cos h + \frac{1}{2}h^2 = 1 + O(h^4).$$ ■

La sección Conjunto de ejercicios 1.3 está disponible en línea. Encuentre la ruta de acceso en las páginas preliminares.

1.4 Software numérico

Los paquetes de software computacional para aproximar las soluciones numéricas a los problemas en muchas formas. En nuestro sitio web para el libro

```
https://sites.google.com/site/numericalanalysis1burden/
```

proporcionamos programas escritos en C, FORTRAN, Maple, Mathematica, MATLAB y Pascal, así como applets de JAVA. Éstos se pueden utilizar para resolver los problemas provistos en los ejemplos y ejercicios, y aportan resultados satisfactorios para muchos de los problemas que usted quizá necesite resolver. Sin embargo, son lo que llamamos programas de *propósito especial*. Nosotros usamos este término para distinguir estos programas de aquellos disponibles en las librerías de subrutinas matemáticas estándar. Los programas en estos paquetes recibirán el nombre de *propósito general*.

Los programas en los paquetes de software de propósito general difieren es sus intenciones de los algoritmos y programas proporcionados en este libro. Los paquetes de software de propósito general consideran formas para reducir los errores debido al redondeo de máquina, el subdesbordamiento y el desbordamiento. También describen el rango de entrada que conducirá a los resultados con cierta precisión específica. Éstas son características dependientes de la máquina, por lo que los paquetes de software de propósito general utilizan parámetros que describen las características de punto flotante de la máquina que se usa para los cálculos.

Ilustración Para ilustrar algunas diferencias entre los programas incluidos en un paquete de propósito general y un programa que nosotros proporcionaríamos en este libro, consideremos un algoritmo que calcula la norma euclidiana de un vector n dimensional $x = (x_1, x_2, \ldots, x_n)^t$. A menudo, esta norma se requiere dentro de los programas más grandes y se define mediante

$$||\mathbf{x}||_2 = \left[\sum_{i=1}^{n} x_i^2 \right]^{1/2}.$$

La norma da una medida de la distancia del vector **x** y el vector **0**. Por ejemplo, el vector $x = (2, 1, 3, 2, 1)^t$ tiene

$$||\mathbf{x}||_2 = [2^2 + 1^2 + 3^2 + (-2)^2 + (-1)^2]^{1/2} = \sqrt{19},$$

por lo que su distancia a partir de $\mathbf{0} = (0, 0, 0, 0, 0)^t$ es $\sqrt{19} \approx 4.36$.

Un algoritmo del tipo que presentaríamos para este problema se proporciona aquí. No incluye parámetros dependientes de máquina y no ofrece garantías de precisión, pero aportará resultados precisos "la mayor parte del tiempo".

ENTRADA $n, x_1, x_2, \ldots, x_n$.

SALIDA $NORM$.

Paso 1 Haga $SUM = 0$.

Paso 2 Para $i = 1, 2, \ldots, n$ determine $SUM = SUM + x_i^2$.

Paso 3 Determine $NORM = SUM^{1/2}$.

Paso 4 SALIDA ($NORM$);
PARE. ■

Un programa con base en nuestro algoritmo es fácil de escribir y comprender. Sin embargo, el programa no sería suficientemente preciso debido a diferentes razones. Por ejemplo, la magnitud de algunos números podría ser demasiado grande o pequeña para representarse con precisión en el sistema de punto flotante de la computadora. Además, este orden para realizar los cálculos quizá no produzca los resultados más precisos, o que la rutina para obtener la raíz cuadrada podría no ser la mejor disponible para el problema. Los diseñadores del algoritmo consideran asuntos de este tipo al escribir programas para software de propósito general. A menudo, estos programas contienen subprogramas para resolver problemas más amplios, por lo que deben incluir controles que nosotros no necesitaremos.

Algoritmos de propósito general

Ahora consideremos un algoritmo para un programa de software de propósito general para calcular la norma euclidiana. Primero, es posible que a pesar de que un componente x_i del vector se encuentre dentro del rango de la máquina, el cuadrado del componente no lo esté. Esto se puede presentar cuando alguna $|x_i|$ es tan pequeña que x_i^2 causa subdesbordamiento o cuando alguna $|x_i|$ es tan grande que x_i^2 causa desbordamiento.

También es posible que todos estos términos se encuentren dentro del rango de la máquina, pero que ocurra desbordamiento a partir de la suma de un cuadrado de uno de los términos para la suma calculada previamente.

Los criterios de precisión dependen de la máquina en la que se realizan los cálculos, por lo que los parámetros dependientes de la máquina se incorporan en el algoritmo. Suponga que trabajamos en una computadora hipotética con base 10, la cual tiene $t \geq 4$ dígitos de precisión, un exponente mínimo *emín* y un exponente máximo *emáx*. Entonces el conjunto de números de punto flotante en esta máquina consiste en 0 y los números de la forma

$$x = f \cdot 10^e, \quad \text{donde} \quad f = \pm(f_1 10^{-1} + f_2 10^{-2} + \cdots + f_t 10^{-t}),$$

donde $1 \leq f_1 \leq 9$ y $0 \leq f_i \leq 9$, para cada $i = 2, \ldots, t$, y donde $emín \leq e \leq emáx$. Estas restricciones implican que el número positivo más pequeño representado en la máquina es $\sigma = 10^{emín-1}$, por lo que cualquier número calculado x con $|x| < \sigma$ causa subdesbordamiento y que x sea 0. El número positivo más grande es $\lambda = (1 - 10^{-t})10^{emáx}$, y cualquier número calculado x con $|x| > \lambda$ causa desbordamiento. Cuando se presenta subdesbordamiento, el programa continuará, a menudo, sin una pérdida significativa de precisión. Si se presenta desbordamiento, el programa fallará.

El algoritmo supone que las características de punto flotante de la máquina se describen a través de los parámetros *N, s, S, y* y *Y*. El número máximo de entradas que se pueden sumar con por lo menos $t/2$ dígitos de precisión está provisto por *N*. Esto implica que el algoritmo procederá a encontrar la norma de un vector $\mathbf{x} = (x_1, x_2, \ldots, x_n)^t$ sólo si $n \leq N$. Para resolver el problema de subdesbordamiento-desbordamiento, los números de punto flotante distintos a cero se dividen en tres grupos:

- números de magnitud pequeña *x*, aquellos que satisfacen $0 < |x| < y$;
- números de magnitud media *x*, donde $y \leq |x| < Y$;
- números de magnitud grande *x*, donde $Y \leq |x|$.

Los parámetros *y* y *Y* se seleccionan con el fin de evitar el problema de subdesbordamiento-desbordamiento al sumar y elevar al cuadrado los números de magnitud media. Elevar al cuadrado los números de magnitud pequeña puede causar subdesbordamiento, por lo que se utiliza un factor de escala *S* mucho mayor a 1 con el resultado $(sx)^2$ que evita el subdesbordamiento incluso cuando x^2 no lo hace. Sumar y elevar al cuadrado los números que tienen una magnitud grande puede causar desbordamiento. Por lo que, en este caso, se utiliza un factor de escala positivo *s* mucho menor a 1 para garantizar que $(sx)^2$ no cause desbordamiento al calcularlo o incluirlo en una suma, a pesar de que x^2 lo haría.

Para evitar escalamiento innecesario, *y* y *Y* se seleccionan de tal forma que el rango de números de magnitud media sea tan largo como sea posible. El siguiente algoritmo es una modificación del que se describe en [Brow, W], p. 471. Éste incluye un procedimiento para sumar los componentes escalados del vector, que son de magnitud pequeña hasta encontrar un componente de magnitud media. Entonces se elimina la escala de la suma previa y continúa al elevar al cuadrado y sumar los números pequeños y medianos hasta encontrar un componente con una magnitud grande. Una vez que el componente con magnitud grande aparece, el algoritmo escala la suma anterior y procede a escalar, elevar al cuadrado y sumar los números restantes.

El algoritmo supone que, en la transición desde números pequeños a medianos, los números pequeños no escalados son despreciables, al compararlos con números medianos. De igual forma, en la transición desde números medianos a grandes, los números medianos no escalados son despreciables, al compararlos con números grandes. Por lo tanto, las selecciones de los parámetros de escalamiento se deben realizar de tal forma que se igualen a 0 sólo cuando son verdaderamente despreciables. Las relaciones comunes entre las características de máquina, como se describen en *t*, σ, λ, *emín* y *emáx* y los parámetros del algoritmo *N, s, S, y* y *Y* se determinan después del algoritmo.

El algoritmo usa tres indicadores para señalar las diferentes etapas en el proceso de suma. Estos indicadores son valores iniciales determinados en el paso 3 del algoritmo. FLAG (BANDERA) 1 es 1 hasta encontrar un componente mediano o grande; entonces se convierte en 0. FLAG (BANDERA) 2 es 0 mientras se suman números pequeños, cambia a 1 cuando se encuentra un número mediano por primera vez, y regresa a 0 cuando se encuentra un número grande. Inicialmente, FLAG (BANDERA) 3 es 0 y cambia a 1 cuando se encuentra un número grande por primera vez. El paso 3 también introduce el indicador DONE (HECHO), que es 0 hasta que se terminan los cálculos y, entonces, regresa a 1.

ENTRADA $N, s, S, y, Y, \lambda, n, x_1, x_2, \ldots, x_n$.

SALIDA *NORM* o un mensaje de error apropiado.

Paso 1 Si $n \leq 0$ entonces SALIDA ('El entero *n* debe ser positivo.')
PARE.

Paso 2 Si $n \geq N$ entonces SALIDA ('El entero *n* es demasiado grande.')
PARE.

Paso 3 Determine $SUM = 0$;
$FLAG1 = 1$; (*Se suman los números pequeños.*)
$FLAG2 = 0$;
$FLAG3 = 0$;
$DONE = 0$;
$i = 1$.

Paso 4 Mientras ($i \le n$ y $FLAG1 = 1$) haga el paso 5.

Paso 5 Si $|x_i| < y$ entonces determine $SUM = SUM + (Sx_i)^2$;
$i = i + 1$
también determine $FLAG1 = 0$. (*Se encuentra un número no pequeño.*)

Paso 6 Si $i > n$ entonces determine $NORM = (SUM)^{1/2}/S$;
$DONE = 1$
también determine $SUM = (SUM/S)/S$; (*Escalamiento de números grandes.*)
$FLAG2 = 1$.

Paso 7 Mientras ($i \le n$ y $FLAG2 = 1$) haga el paso 8. (*Se suman los números medianos.*)

Paso 8 Si $|x_i| < Y$ entonces determine $SUM = SUM + x_i^2$;
$i = i + 1$
también determine $FLAG2 = 0$. (*Se encuentra un número no grande.*)

Paso 9 Si $DONE = 0$ entonces
si $i > n$ entonces determine $NORM = (SUM)^{1/2}$;
$DONE = 1$
también determine $SUM = ((SUM)s)s$; (*Escalamiento de números grandes.*)
$FLAG3 = 1$.

Paso 10 Mientras $i \le n$ y $FLAG3 = 1$) haga el paso 11.

Paso 11 Determine $SUM = SUM + (sx_i)^2$; (*Sume los números grandes.*)
$i = i + 1$.

Paso 12 Si $DONE = 0$ entonces
si $SUM^{1/2} < \lambda s$ entonces determine $NORM = (SUM)^{1/2}/s$;
$DONE = 1$
también determine $SUM = \lambda$. (*La norma es demasiado grande.*)

Paso 13 Si $DONE = 1$ entonces SALIDA (‘Norma es’, $NORM$)
también SALIDA (‘Norma $\ge$’, $NORM$, 'ocurrió sobreflujo').

Paso 14 PARE.

Las relaciones entre las características de máquina t, σ, λ, *emín* y *emáx* y los parámetros del algoritmo N, s, S, y y Y se seleccionaron en [Brow, W], p. 471, como

$N = 10^{e_N}$, donde $e_N = \lfloor (t-2)/2 \rfloor$, El entero más grande menor o igual a $(t-2)/2$;

$s = 10^{e_s}$, donde $e_s = \lfloor -(emáx + e_N)/2 \rfloor$;

$S = 10^{e_S}$, donde $e_S = \lceil (1 - emín)/2 \rceil$, El entero más pequeño mayor o igual que $(1 - emín)/2$;

$y = 10^{e_y}$, donde $e_y = \lceil (emín + t - 2)/2 \rceil$;

$Y = 10^{e_Y}$, donde $e_Y = \lfloor (emáx - e_N)/2 \rfloor$.

La primera computadora portátil fue la Osborne I, producida en 1981, a pesar de que era mucho más grande y pesada de lo que podríamos pensar como portátil.

El sistema FORTRAN (FORmula TRANSlator) fue el lenguaje de programación científica de propósito general original. Sigue utilizándose ampliamente en situaciones que requieren cálculos científicos intensivos.

El proyecto EISPACK fue el primer paquete de software numérico a gran escala en estar disponible para dominio público y lideró el camino para que muchos paquetes lo siguieran.

La fiabilidad construida en este algoritmo ha incrementado ampliamente la complejidad, en comparación con el algoritmo provisto antes en esta sección. En la mayoría de los casos, los algoritmos de propósito especial y general proporcionan resultados idénticos. La ventaja del algoritmo de propósito general es que proporciona seguridad para sus resultados.

Existen muchas formas de software numérico de propósito general disponibles en el ámbito comercial y en el dominio público. La mayor parte de los primeros se escribió para las computadoras centrales, y una buena referencia es *Sources and Development of Mathematical Software (Fuentes y desarrollo de software matemático)*, editado por Wayne Cowell [Co].

Ahora que las computadoras personales son suficientemente poderosas, existe software numérico estándar para ellas. La mayoría de este software numérico se escribe en FORTRAN, a pesar de que algunos están escritos en C, C++ y FORTRAN90.

Los procedimientos ALGOL se presentaron para cálculos de matrices en 1971 en [WR]. Después, se desarrolló un paquete de subrutinas FORTRAN con base principalmente en los procedimientos ALGOL dentro de las rutinas EISPACK. Estas rutinas se documentan en los manuales publicados por Springer-Verlag como parte de sus *Lecture Notes (Notas de clase)* en la serie *Computer Science (Ciencias computacionales)* [Sm, B] y [Gar]. Las subrutinas FORTRAN se utilizan para calcular los valores propios y los vectores propios para una variedad de diferentes tipos de matrices.

LINPACK es un paquete de subrutinas FORTRAN para analizar y resolver sistemas de ecuaciones lineales y resolver problemas de mínimos cuadrados lineales. La documentación en este paquete se encuentra en [DBMS]. Una introducción paso a paso para LINPACK, EISPACK y BLAS (Basic Linear Algebra Subprograms; Subprogramas de Álgebra Lineal Básica) se proporciona en [CV].

El paquete LAPACK, disponible por primera vez en 1992, es una librería de las subrutinas FORTRAN que sustituyen a LINPACK y EISPACK al integrar estos dos conjuntos de algoritmos en un paquete unificado y actualizado. El software se ha reestructurado para lograr mayor eficiencia en procesadores de vectores y otros multiprocesadores de alto desempeño y memoria compartida. LAPACK se expande en profundidad y amplitud en la versión 3.0, disponible en FORTRAN, FORTRAN90, C, C++ y JAVA. C y JAVA sólo están disponibles como interfaces de idioma o traducciones de las librerías FORTRAN de LAPACK. El paquete BLAS no forma parte de LAPACK, pero el código para BLAS se distribuye con LAPACK

Otros paquetes para resolver tipos específicos de problemas están disponibles en el dominio público. Como alternativa para netlib, puede utilizar Xnetlib para buscar en la base de datos y recuperar software. Encuentre más información en el artículo *Software Distribution Using Netlib* de Dongarra Roman y Wade [DRW].

La ingeniería de software se estableció como disciplina de laboratorio durante las décadas de 1970 y 1980. EISPACK se desarrolló en Argonne Labs y LINPACK poco después. A principios de la década de 1980, Argonne fue reconocido en el ámbito internacional como líder mundial en cálculos simbólicos y numéricos.

Estos paquetes de software son muy eficientes, precisos y confiables. Se prueban de manera meticulosa y la documentación está disponible fácilmente. A pesar de que los paquetes son portátiles, es una buena idea investigar la dependencia de la máquina y leer la documentación con todo detalle. Los programas prueban casi todas las contingencias especiales que podrían resultar en errores o fallas. Al final de cada capítulo analizaremos algunos de los paquetes de propósito general adecuados.

Los paquetes comercialmente disponibles también representan los métodos numéricos de vanguardia. A menudo, su contenido está basado en los paquetes de dominio público, pero incluyen métodos en bibliotecas para casi cualquier tipo de problemas.

Las IMSL (International Mathematical and Statistical Libraries; Bibliotecas Estadísticas y Matemáticas Internacionales) están formadas por bibliotecas MATH, STAT y SFUN para matemáticas numéricas, estadísticas y funciones especiales, respectivamente. Estas bibliotecas contienen más de 900 subrutinas originalmente disponibles en FOR-TRAN 77 y ahora disponibles en C, FORTRAN90 y JAVA. Estas subrutinas resuelven los problemas de análisis numéricos más comunes. Las librerías están comercialmente disponibles en Visual Numerics.

En 1970, IMSL se convirtió en la primera librería científica a gran escala para computadoras centrales. Ya que en esa época, existían librerías para sistemas computacionales que iban desde supercomputadoras hasta computadoras personales.

Los paquetes se entregan en formato compilado con documentación amplia. Existe un programa de ejemplo para cada rutina, así como información de referencia de fondo. IMSL contiene métodos para sistemas lineales, análisis de sistemas propios, interpolación y aproximación, integración y diferenciación, ecuaciones diferenciales, transformadas, ecuaciones no lineales, optimización y operaciones básicas matriz/vector. La biblioteca también incluye amplias rutinas estadísticas.

El Numerical Algorithms Group (NAG) se fundó en Reino Unido en 1971 y desarrolló la primera biblioteca de software matemático. Actualmente tiene más de 10 000 usuarios a nivel mundial y contiene más de 1000 funciones matemáticas y estadísticas que van desde software de simulación estadística, simbólica, visualización y numérica hasta compiladores y herramientas de desarrollo de aplicaciones.

El Numerical Algorithms Group (NAG; Grupo de Algoritmos Numéricos) ha existido en Reino Unido desde 1970. NAG ofrece más de 1000 subrutinas en una biblioteca FORTRAN 77, aproximadamente 400 subrutinas en una biblioteca C, más de 200 subrutinas en una biblioteca FORTRAN 90 y una biblioteca MPI FORTRAN para máquinas paralelas y agrupaciones de estaciones de trabajo o computadoras personales. Una introducción útil para las rutinas NAG es [Ph]. La biblioteca NAG contiene rutinas para realizar la mayor parte de las tareas de análisis numérico estándar de manera similar a la de IMSL. También incluye algunas rutinas estadísticas y un conjunto de rutinas gráficas.

Los paquetes IMSL y NAG están diseñados para los matemáticos, científicos o ingenieros que desean llamar subrutinas de alta calidad de C, Java o FORTRAN desde dentro de un programa. La documentación disponible con los paquetes comerciales ilustra el programa activador común, requerido para utilizar las rutinas de la librería. Los siguientes tres paquetes de software son ambientes autónomos. Cuando se activan, los usuarios introducen comandos para hacer que el paquete resuelva un problema. Sin embargo, cada paquete permite programar dentro del lenguaje de comando.

Originalmente, MATLAB se escribió para proporcionar acceso al software de matriz desarrollado en proyectos LINPACK y EISPACK. La primera versión se escribió a finales de la década de 1970 para utilizarse en cursos de teoría de matriz, álgebra lineal y análisis numérico. Actualmente, existen más de 500 000 usuarios de MATLAB en más de 100 países.

MATLAB es una matriz de laboratorio que, originalmente, era un programa Fortran publicado por Cleve Moler [Mo] en la década de 1980. El laboratorio está basado principalmente en las subrutinas EISPACK y LINPACK, a pesar de que se han integrado funciones, como sistemas no lineales, integración numérica, splines cúbicos, ajuste de curvas, optimización, ecuaciones diferenciales normales y herramientas gráficas. Actualmente, MATLAB está escrito en C y lenguaje ensamblador y la versión para PC de este paquete requiere un coprocesador numérico. La estructura básica es realizar operaciones de matriz, como encontrar los valores propios de una matriz introducida desde la línea de comando o desde un archivo externo a través de llamadas a funciones. Éste es un sistema autónomo poderoso que es especialmente útil para instrucción en un curso de álgebra lineal aplicada.

El segundo paquete es GAUSS, un sistema matemático y estadístico producido por Lee E. Ediefson y Samuel D. Jones en 1985. Está codificado principalmente en lenguaje ensamblador y basado EISPACK y LINPACK. Al igual que en el caso de MATLAB, existe integración/diferenciación, sistemas no lineales, transformadas rápidas de Fourier y gráficas. GAUSS se orienta menos hacia la enseñanza de álgebra lineal y más hacia el análisis estadístico de datos. Este paquete también utiliza un coprocesador numérico, cuando existe uno.

Las rutinas NAG son compatibles con Maple, desde la versión 9.0.

El tercer paquete es Maple, un sistema de álgebra computacional desarrollado en 1980 Mediante el Symbolic Computational Group (Grupo Computacional Simbólico) en la University of Waterloo. El diseño para el sistema original Maple se presenta en el artículo de B.W. Char, K.O. Geddes, W.M. Gentlemen y G.H. Gonnet [CGGG].

Maple, escrito en C, tiene la capacidad de manipular información de manera simbólica. Esta manipulación simbólica permite al usuario obtener respuestas exactas, en lugar de valores numéricos. Maple puede proporcionar respuestas exactas a problemas matemáticos como integrales, ecuaciones diferenciales y sistemas lineales. Contiene una estructura de programación y permite texto, así como comandos, para guardarlos en sus archivos de hojas de trabajo. Estas hojas de trabajo se pueden cargar a Maple y ejecutar los comandos.

El igualmente popular Mathematica, liberado en 1988, es similar a Maple.

Existen numerosos paquetes que se pueden clasificar como paquetes de supercalculadora para la PC. Sin embargo, éstos no deberían confundirse con el software de propósito general que se ha mencionado aquí. Si está interesado en uno de estos paquetes, debería leer *Supercalculators on the PC (Supercalculadoras en la PC)* de B. Simon y R. M. Wilson [SW].

Información adicional sobre el software y las bibliotecas de software se puede encontrar en los libros de Cody y Waite [CW] y de Kockler [Ko] y el artículo de 1995 de Dongarra y Walker [DW]. Más información sobre cálculo de punto flotante se puede encontrar en el libro de Chaitini-Chatelin y Frayse [CF] y el artículo de Goldberg [Go].

Los libros que abordan la aplicación de técnicas numéricas sobre computadoras paralelas incluyen los de Schendell [Sche], Phillips and Freeman [PF] y Golub y Ortega [GO].

Las secciones Pregunta de análisis, Conceptos clave y Revisión del capítulo están disponibles en línea. Encuentre la ruta de acceso en las páginas preliminares.

CAPÍTULO

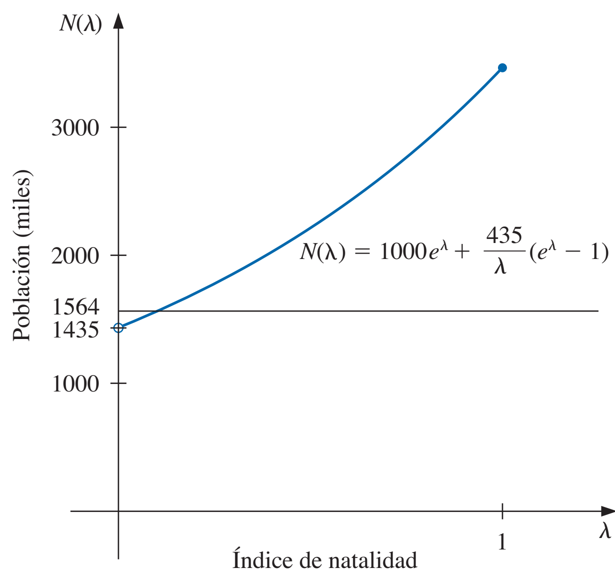
2 Soluciones de las ecuaciones en una variable

Introducción

A menudo, el crecimiento de una población se puede modelar sobre periodos breves al asumir que aumenta de manera continua con el tiempo a una tasa proporcional al número actual en ese momento. Suponga que $N(t)$ denota el número en la población en el tiempo t y λ denota la tasa constante de natalidad. Entonces, dicha población satisface la ecuación diferencial

$$\frac{dN(t)}{dt} = \lambda N(t),$$

cuya solución es $N(t) = N_0 e^{\lambda t}$ donde N_0 denota la población inicial.



Este modelo exponencial sólo es válido cuando la población está aislada, sin inmigración. Si se permite inmigración a una tasa constante v, entonces la ecuación diferencial se convierte en

$$\frac{dN(t)}{dt} = \lambda N(t) + v,$$

cuya solución es

$$N(t) = N_0 e^{\lambda t} + \frac{v}{\lambda}(e^{\lambda t} - 1).$$

Suponga que, en un inicio, cierta población contiene $N(0) = 1\,000\,000$ individuos, que 435 000 individuos inmigran a la comunidad durante el primer año y que existen $N(1) = 1\,564\,000$ individuos al final del año. Para determinar la natalidad de esta población, necesitamos encontrar λ en la ecuación

$$1\,564\,000 = 1\,000\,000e^{\lambda} + \frac{435\,000}{\lambda}(e^{\lambda} - 1).$$

En esta ecuación no es posible resolver de manera explícita para λ, pero los métodos numéricos que se analizan en este capítulo se pueden utilizar para aproximar las soluciones de las ecuaciones de este tipo con una precisión arbitrariamente alta. Las solución a este problema particular se considera en el ejercicio 22 de la sección 2.3.

2.1 El método de bisección

En este capítulo consideramos uno de los problemas básicos de la aproximación numérica, el **problema de la búsqueda de la raíz**. Este proceso implica encontrar una **raíz**, o solución, para una ecuación de la forma $f(x) = 0$, para una función f dada. Una raíz de esta ecuación también recibe el nombre de **cero** de la función f.

El problema de encontrar una aproximación para la raíz de una ecuación se puede rastrear por lo menos al año 1 700 a.C. Una tabla cuneiforme en la Colección Babilónica de Yale que data del periodo provee un número sexagesimal (base 60) equivalente a 1.414222 como una aproximación para $\sqrt{2}$, un resultado que es preciso dentro de 10^{-5}. Esta aproximación se puede encontrar al aplicar una técnica descrita en el ejercicio 19 de la sección 2.2.

Técnica de bisección

En las ciencias computacionales, al proceso que consiste en dividir a la mitad un conjunto de manera continua para buscar la solución de un problema, como lo hace el método de bisección, se le conoce como procedimiento de *búsqueda binaria*.

La primera técnica, basada en el teorema de valor intermedio, recibe el nombre de **bisección,** o método de búsqueda binaria.

Suponga que f es una función continua definida dentro del intervalo $[a, b]$ con $f(a)$ y $f(b)$ de signos opuestos. El teorema de valor intermedio implica que existe un número p en (a, b) con $f(p) = 0$. A pesar de que el procedimiento operará cuando haya más de una raíz en el intervalo (a, b), para simplicidad, nosotros asumimos que la raíz en este intervalo es única. El método realiza repetidamente una reducción a la mitad (o bisección) de los subintervalos de $[a, b]$ y, en cada paso, localizar la mitad que contiene p.

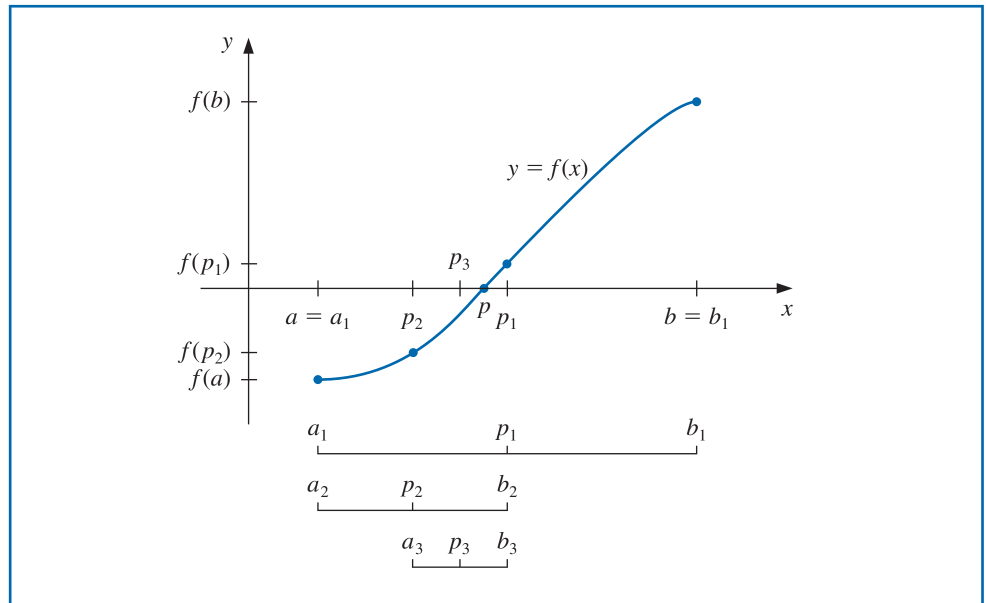
Para comenzar, sea $a_1 = a$ y $b_1 = b$ y sea p_1 es el punto medio de $[a, b]$, es decir,

$$p_1 = a_1 + \frac{b_1 - a_1}{2} = \frac{a_1 + b_1}{2}.$$

- Si $f(p_1) = 0$, entonces $p = p_1$ y terminamos.
- Si $f(p_1) \neq 0$, entonces $f(p_1)$ tiene el mismo signo que ya sea $f(a_1)$ o $f(b_1)$.
    - ◇ Si $f(p_1)$ y $f(a_1)$ tienen el mismo signo, $p \in (p_1, b_1)$. Sea $a_2 = p_1$ y $b_2 = b_1$.
    - ◇ Si $f(p_1)$ y $f(a_1)$ tienen signos opuestos, $p \in (a_1, p_1)$. Sea $a_2 = a_1$ y $b_2 = p_1$.

Entonces, volvemos a aplicar el proceso al intervalo $[a_2, b_2]$. Esto produce el método descrito en el algoritmo 2.1 (consulte la figura 2.1).

Figura 2.1



ALGORITMO **2.1**

Bisección

Para encontrar una solución a $f(x) = 0$ dada la función continua determinada f en el intervalo $[a, b]$, donde $f(a)$ y $f(b)$ tienen signos opuestos:

ENTRADA puntos finales a, b; tolerancia TOL; número máximo de iteraciones N_0.

SALIDA solución aproximada p o mensaje de falla.

Paso 1 Sea $i = 1$;
$FA = f(a)$.

Paso 2 Mientras $i \leq N_0$ haga los pasos 3–6.

Paso 3 Sea $p = a + (b - a)/2$; (*Calcule* p_i.)
$FP = f(p)$.

Paso 4 Si $FP = 0$ o $(b - a)/2 < TOL$ entonces
SALIDA (p); (*Procedimiento completado exitosamente*.)
PARE.

Paso 5 Sea $i = i + 1$.

Paso 6 Si $FA \cdot FP > 0$ entonces determine $a = p$; (*Calcule* a_i, b_i.)
$FA = FP$
también determine $b = p$. (*FA no cambia*.)

Paso 7 SALIDA ('El método fracasó después de N_0 iteraciones, N_0 =', N_0);
(*El procedimiento no fue exitoso*.)
PARE.

■

Se pueden aplicar otros procedimientos de parada en el paso 4 del algoritmo 2.1 o en cualquier técnica iterativa en este capítulo. Por ejemplo, podemos seleccionar una tolerancia $\epsilon > 0$ y generar $p_1, \ldots, p_N$ hasta que se cumpla una de las siguientes condiciones:

$$|p_N - p_{N-1}| < \epsilon, \tag{2.1}$$

$$\frac{|p_N - p_{N-1}|}{|p_N|} < \epsilon, \quad p_N \neq 0, \quad \text{o} \tag{2.2}$$

$$|f(p_N)| < \epsilon. \tag{2.3}$$

Por desgracia, se pueden presentar dificultades al utilizar cualquiera de estos criterios de parada. Por ejemplo, existen sucesiones $\{p_n\}_{n=0}^{\infty}$ con la propiedad de que las diferencias $p_n - p_{n-1}$ convergen a cero mientras la sucesión misma diverge (consulte el ejercicio 19). También es posible que $f(p_n)$ esté cerca de cero mientras p_n difiere significativamente de p (consulte el ejercicio 20). Sin conocimiento adicional sobre f o p, la desigualdad (2.2) es el mejor criterio de parada que se puede aplicar porque verifica el error relativo.

Al utilizar una computadora para generar aproximaciones, es bueno practicar la configuración de un límite superior sobre el número de iteraciones. Esto elimina la posibilidad de entrar en un ciclo infinito, una situación que puede surgir cuando la sucesión diverge (y también cuando el programa se codifica de manera incorrecta). Eso se realizó en el paso 2 del algoritmo 2.1, donde se estableció el límite N_0 y se concluyó el procedimiento si $i > N_0$.

Observe que para iniciar el algoritmo de bisección, se debe encontrar un intervalo $[a, b]$ con $f(a) \cdot f(b) < 0$. En cada paso, la longitud del intervalo conocida por contener un cero de f se reduce en un factor de 2; por lo tanto, es ventajoso elegir el intervalo inicial $[a, b]$ tan pequeño como sea posible. Por ejemplo, si $f(x) = 2x^3 - x^3 + x - 1$, tenemos

$$f(-4) \cdot f(4) < 0 \quad \text{y} \quad f(0) \cdot f(1) < 0,$$

por lo que el algoritmo de bisección se puede usar en $[-4, 4]$ o en $[0, 1]$. Al iniciar el algoritmo de bisección en $[0, 1]$ en lugar de $[-4, 4]$ reduciremos en 3 el número de iteraciones requerido para alcanzar una precisión específica.

El siguiente ejemplo ilustra el algoritmo de bisección. La iteración en este ejemplo termina cuando una cota para el error relativo es menor a 0.0001. Esto se garantiza al tener

$$\frac{|p - p_n|}{\text{mín}\{|a_n|, |b_n|\}} < 10^{-4}.$$

Ejemplo 1 Muestre que $f(x) = x^3 + 4x^2 - 10 = 0$ tiene una raíz en $[1, 2]$ y utilice el método de bisección para determinar una aproximación para la raíz que sea precisa por lo menos dentro de 10^{-4}.

Solución Puesto que $f(1) = -5$ y $f(2) = 14$, el teorema de valor intermedio 1.11 garantiza que esta función continua tenga una raíz en $[1, 2]$.

Para la primera iteración del método de bisección, usamos el hecho de que en el punto medio de $[1, 2]$ tenemos $f(1.5) = 2.375 > 0$. Esto indica que deberíamos seleccionar el intervalo $[1, 1.5]$ para nuestra segunda iteración. A continuación, encontramos que $f(1.25) = -1.796875$, por lo que nuestro intervalo se vuelve $[1.25, 1.5]$, cuyo punto medio es 1.375. Si continuamos de esta forma, obtenemos los valores en la tabla 2.1.

Después de 13 iteraciones, $p_{13} = 1.365112305$ se aproxima a la raíz p con un error

$$|p - p_{13}| < |b_{14} - a_{14}| = |1.365234375 - 1.365112305| = 0.000122070.$$

Ya que $|a_{14}| < |p|$, tenemos

$$\frac{|p - p_{13}|}{|p|} < \frac{|b_{14} - a_{14}|}{|a_{14}|} \leq 9.0 \times 10^{-5},$$

Tabla 2.1

n	a_n	b_n	p_n	$f(p_n)$
1	1.0	2.0	1.5	2.375
2	1.0	1.5	1.25	−1.79687
3	1.25	1.5	1.375	0.16211
4	1.25	1.375	1.3125	−0.84839
5	1.3125	1.375	1.34375	−0.35098
6	1.34375	1.375	1.359375	−0.09641
7	1.359375	1.375	1.3671875	0.03236
8	1.359375	1.3671875	1.36328125	−0.03215
9	1.36328125	1.3671875	1.365234375	0.000072
10	1.36328125	1.365234375	1.364257813	−0.01605
11	1.364257813	1.365234375	1.364746094	−0.00799
12	1.364746094	1.365234375	1.364990235	−0.00396
13	1.364990235	1.365234375	1.365112305	−0.00194

de modo que la aproximación es correcta por lo menos dentro de 10^{-4}. El valor correcto de p para nueve lugares decimales es $p = 1.365230013$. Observe que p_9 está más cerca de p que es la aproximación final p_{13}. Se podría sospechar que esto es verdad porque $|f(p_9)| < |f(p_{13})|$, pero no podemos asegurarlo a menos que conozcamos la respuesta verdadera. ■

El método de bisección, a pesar de que está conceptualmente claro, tiene desventajas significativas. Su velocidad de convergencia es más lenta (es decir, N se puede volver bastante grande antes de que $|p - p_N|$ sea suficientemente pequeña) y se podría descartar inadvertidamente una buena aproximación intermedia. Sin embargo, el método tiene la importante propiedad de que siempre converge a una solución y por esta razón con frecuencia se utiliza como iniciador para los métodos más eficientes que veremos más adelante en este capítulo.

Teorema 2.1 Suponga que $f \in C[a, b]$ y $f(a) \cdot f(b) < 0$. El método de bisección genera una sucesión $\{p_n\}_{n=1}^{\infty}$ que se aproxima a cero p de f con

$$|p_n - p| \leq \frac{b-a}{2^n}, \quad \text{cuando} \quad n \geq 1.$$

Demostración Para cada $n \geq 1$, se tiene

$$b_n - a_n = \frac{1}{2^{n-1}}(b-a) \quad \text{y} \quad p \in (a_n, b_n).$$

Ya que $p_n = \frac{1}{2}(a_n + b_n)$ para toda $n \geq 1$, se sigue que

$$|p_n - p| \leq \frac{1}{2}(b_n - a_n) = \frac{b-a}{2^n}.$$ ■

porque

$$|p_n - p| \leq (b-a)\frac{1}{2^n},$$

la sucesión $\{p_n\}_{n=1}^{\infty}$ converge en p con una razón de convergencia $O\left(\frac{1}{2^n}\right)$; es decir

$$p_n = p + O\left(\frac{1}{2^n}\right).$$

Es importante señalar que el teorema 2.1 sólo provee una cota para el error de aproximación y que ésta podría ser bastante conservadora. Por ejemplo, cuando se aplica al problema

en el ejemplo 1 sólo garantiza que

$$|p - p_9| \leq \frac{2-1}{2^9} \approx 2 \times 10^{-3},$$

pero el error real es mucho menor:

$$|p - p_9| = |1.365230013 - 1.365234375| \approx 4.4 \times 10^{-6}.$$

Ejemplo 2 Determine el número de iteraciones necesarias para resolver $f(x) = x^3 + 4x^2 - 10 = 0$ con precisión de 10^{-3} mediante $a_1 = 1$ y $b_1 = 2$.

Solución Usaremos logaritmos para encontrar un entero N que satisface

$$|p_N - p| \leq 2^{-N}(b-a) = 2^{-N} < 10^{-3}.$$

Los logaritmos para cualquier base bastarían, pero usaremos logaritmos base 10 porque la tolerancia se determina como una potencia de 10. Ya que $2^{-N} < 10^{-3}$ implica que $\log_{10} 2^{-N} < \log_{10} 10^{-3} = -3$, tenemos

$$-N \log_{10} 2 < -3 \quad \text{y} \quad N > \frac{3}{\log_{10} 2} \approx 9.96.$$

Por lo tanto, se requieren 10 iteraciones para una aproximación precisa dentro de 10^{-3}.

La tabla 2.1 muestra que el valor de $p_9 = 1.365234375$ es exacto dentro de 10^{-4}. De nuevo, es importante considerar que el análisis de error sólo proporciona una cota para el número de iteraciones. En muchos casos, la cota es mucho mayor que el número real requerido. ■

La cota para el número de iteraciones en el método de bisección supone que los cálculos se realizan con aritmética de dígitos infinitos. Al implementar el método en una computadora, necesitamos considerar los efectos del error de redondeo. Por ejemplo, el cálculo del punto medio del intervalo $[a_n, b_n]$ se debería encontrar a partir de la ecuación

$$p_n = a_n + \frac{b_n - a_n}{2} \quad \text{en lugar de} \quad p_n = \frac{a_n + b_n}{2}.$$

La primera ecuación añade una pequeña corrección, $(b_n - a_n)/2$, al valor conocido a_n. Cuando $b_n - a_n$ está cerca de la precisión máxima de la máquina, esta corrección podría ser un error, pero el error no afectaría significativamente el valor calculado de p_n. Sin embargo, es posible que $(b_n = a_n)/2$ regrese a un punto medio que no se encuentra dentro del intervalo $[a_n, b_n]$.

La palabra en latín *signum* significa "señal" o "signo". Por lo que la función signum retorna de forma bastante natural el signo de un número (a menos que el número sea cero).

Como observación final, para determinar el subintervalo de $[a_n, b_n]$ contiene una raíz de f, es mejor utilizar la función **signum**, que se define como

$$\text{sgn}(x) = \begin{cases} -1, & \text{si } x < 0, \\ 0, & \text{si } x = 0, \\ 1, & \text{si } x > 0. \end{cases}$$

Utilizar

$$\text{sgn}\,(f(a_n))\,\text{sgn}\,(f(p_n)) < 0 \qquad \text{en lugar de} \qquad f(a_n)f(p_n) < 0$$

nos da el mismo resultado, pero evita la posibilidad de desbordamiento o subdesbordamiento en la multiplicación de $f(a_n)$ y $f(p_n)$.

La sección Conjunto de ejercicios 2.1 está disponible en línea. Encuentre la ruta de acceso en las páginas preliminares.



2.2 Iteración de punto fijo

Un *punto fijo* para una función es un número en el que el valor de la función no cambia cuando se aplica la función.

Definición 2.2 El número p es un **punto fijo** para una función dada g si $g(p) = p$. ■

En esta sección, consideramos el problema de encontrar soluciones para los problemas de punto fijo y la conexión entre los problemas de punto fijo y aquellos para encontrar la raíz que queremos resolver. Los problemas para encontrar la raíz y los de punto fijo son clases equivalentes en el siguiente sentido:

Los resultados de punto fijo ocurren en muchas áreas de las matemáticas y son una herramienta principal de los economistas para proporcionar resultados concernientes a los equilibrios. Aunque la idea detrás de la técnica es antigua, la terminología fue usada primero por el matemático holandés L. E. J. Brouwer (1882-1962) a principios de 1900.

- Dado un problema para encontrar la raíz $f(p) = 0$, podemos definir las funciones g con un punto fijo en p en diferentes formas, por ejemplo,

$$g(x) = x - f(x) \quad \text{o} \quad g(x) = x + 3f(x).$$

- Por otra parte, si la función g tiene un punto fijo en p, entonces la función definida por

$$f(x) = x - g(x)$$

  tiene un cero en p.

A pesar de que los problemas que queremos resolver se encuentran en la forma para encontrar la raíz, la forma de punto fijo es más fácil de analizar y ciertas elecciones de punto fijo conducen a técnicas muy poderosas para encontrar la raíz.

Primero necesitamos estar cómodos con este nuevo tipo de problema y decidir cuándo una función tiene un punto fijo y cómo se pueden aproximar los puntos fijos dentro de una precisión especificada.

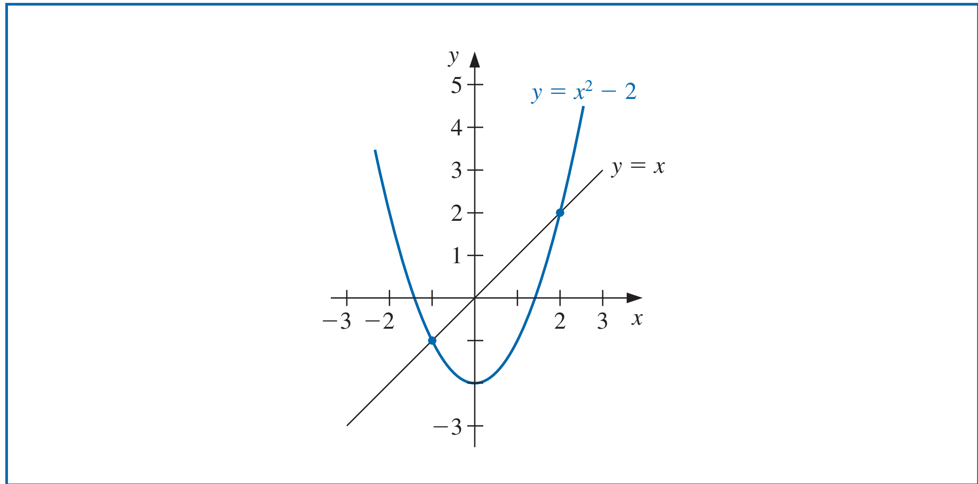
Ejemplo 1 Determine cualquier punto fijo de la función $g(x) = x^2 - 2$.

Solución Un punto fijo p para g tiene la propiedad de que

$$p = g(p) = p^2 - 2, \quad \text{lo cual implica que} \quad 0 = p^2 - p - 2 = (p + 1)(p - 2).$$

Un punto fijo para g ocurre precisamente cuando la gráfica de $y = g(x)$ interseca la gráfica de $y = x$, por lo que g tiene dos puntos fijos, uno en $p = -1$ y el otro en $p = 2$. Éstos se muestran en la figura 2.2. ■

Figura 2.2



El siguiente teorema proporciona suficientes condiciones para la existencia y unicidad de un punto fijo.

Teorema 2.3

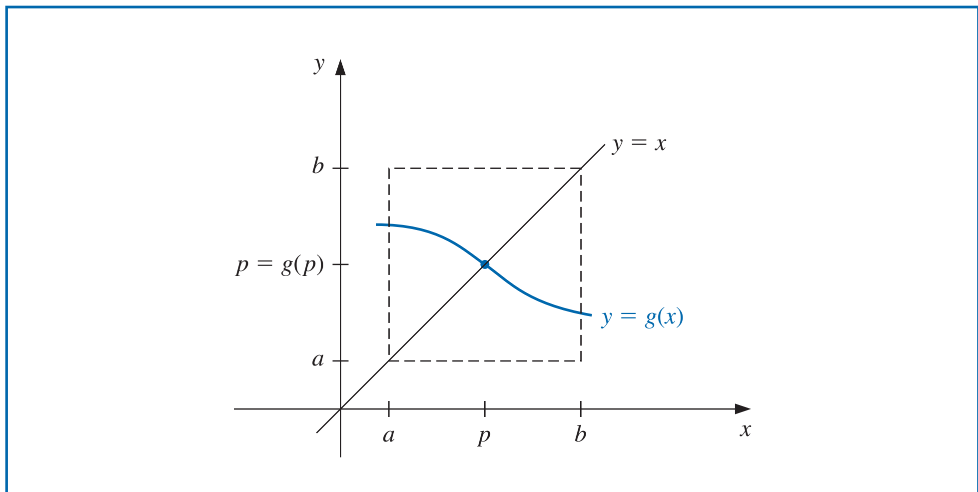
i) Si $g \in C[a, b]$ y $g(x) \in [a, b]$ para todas $x \in [a, b]$, entonces g tiene por lo menos un punto fijo en $[a, b]$.

ii) Si, además, $g'(x)$ existe en (a, b) y hay una constante positiva $k < 1$ con

$$|g'(x)| \leq k, \quad \text{para todas las } x \in (a, b),$$

entonces, existe exactamente un punto fijo en $[a, b]$. (Véase la figura 2.3)

Figura 2.3



Demostración

i) Si $g(a) = a$ o $g(b) = b$, entonces g tiene un punto fijo en un extremo. De lo contrario, entonces $g(a) > a$ y $g(b) < b$. La función $h(x) = g(x) - x$ es continua en $[a, b]$, con

$$h(a) = g(a) - a > 0 \quad \text{y} \quad h(b) = g(b) - b < 0.$$

El teorema de valor intermedio implica que existe $p \in (a, b)$ para la cual $h(p) = 0$. Este número p es un punto fijo para g porque

$$0 = h(p) = g(p) - p \quad \text{implica que} \quad g(p) = p.$$

ii) Suponga, además, que $|g'(x)| \le k < 1$ y que p y q son puntos fijos en $[a, b]$. Si $p \neq q$, entonces el teorema de valor medio implica que existe un número ξ entre p y q y por lo tanto en $[a, b]$ con

$$\frac{g(p) - g(q)}{p - q} = g'(\xi).$$

Por lo tanto

$$|p - q| = |g(p) - g(q)| = |g'(\xi)||p - q| \le k|p - q| < |p - q|,$$

lo cual es una contradicción. Esta contradicción debe provenir de la única suposición $p \neq q$. Por lo tanto, $p = q$ y el punto fijo en $[a, b]$ es único. ■

Ejemplo 2 Muestre que $g(x) = (x^2 - 1)/3$ tiene un punto fijo único en el intervalo $[-1, 1]$.

Solución Los valores máximo y mínimo de $g(x)$ para x en $[-1, 1]$ deben ocurrir ya sea cuando x es un extremo del intervalo o cuando la derivada es 0. Puesto que $g'(x) = 2x/3$, la función g es continua y $g'(x)$ existe en $[-1, 1]$. Los valores máximo y mínimo de $g(x)$ se presentan en $x = -1$, $x = 0$ o $x = 1$. Pero $g(-1) = 0$, $g(1) = 0$ y $g(0) = -1/3$, por lo que un máximo absoluto para $g(x)$ en $[-1, 1]$ se presenta en $x = -1$ y $x = 1$ y un mínimo absoluto en $x = 0$.

Además,

$$|g'(x)| = \left|\frac{2x}{3}\right| \le \frac{2}{3}, \quad \text{para todas las} \quad x \in (-1, 1).$$

De este modo g satisface todas las hipótesis del teorema 2.3 y tiene un punto fijo único en $[-1, 1]$. ■

Para la función en el ejemplo 2, el único punto fijo p en el intervalo $[-1, 1]$ se puede determinar de manera algebraica. Si

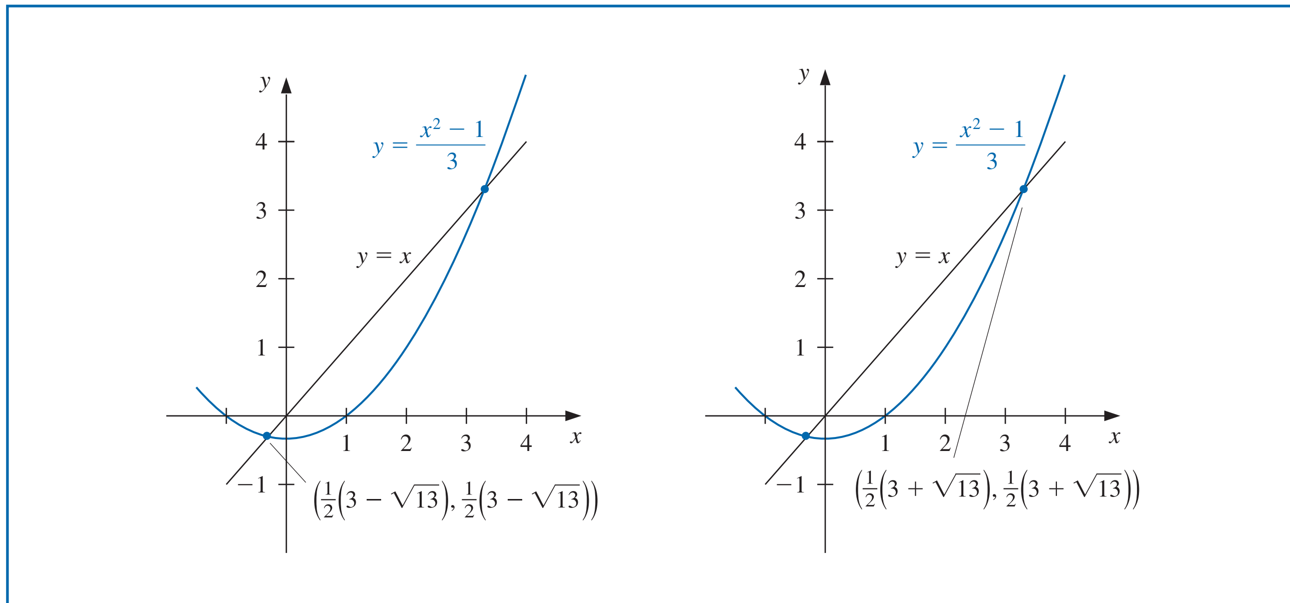
$$p = g(p) = \frac{p^2 - 1}{3}, \quad \text{entonces} \quad p^2 - 3p - 1 = 0,$$

lo cual, por la fórmula cuadrática, implica, como se muestra en la gráfica izquierda en la figura 2.4, que

$$p = \frac{1}{2}(3 - \sqrt{13}).$$

Observe que g también tiene un punto fijo único $p = \frac{1}{2}(3+\sqrt{13})$ para el intervalo $[3, 4]$. Sin embargo, $g(4) = 5$ y $g'(4) = \frac{8}{3} > 1$, por lo que g no satisface la hipótesis del teorema 2.3 en $[3, 4]$. Esto demuestra que la hipótesis del teorema 2.3 es suficiente, pero no necesaria para garantizar la unicidad del punto fijo. (Consulte la gráfica de la derecha en la figura 2.4.)

Figura 2.4



Ejemplo 3 Muestre que el teorema 2.3 no garantiza un punto fijo único de $g(x) = 3^{-x}$ en el intervalo $[0, 1]$, aunque existe un punto fijo único en este intervalo.

Solución $g'(x) = -3^{-x} \ln 3 < 0$ en $[0, 1]$, la función g es estrictamente decreciente en $[0, 1]$. Por lo que

$$g(1) = \frac{1}{3} \leq g(x) \leq 1 = g(0), \quad \text{para} \quad 0 \leq x \leq 1.$$

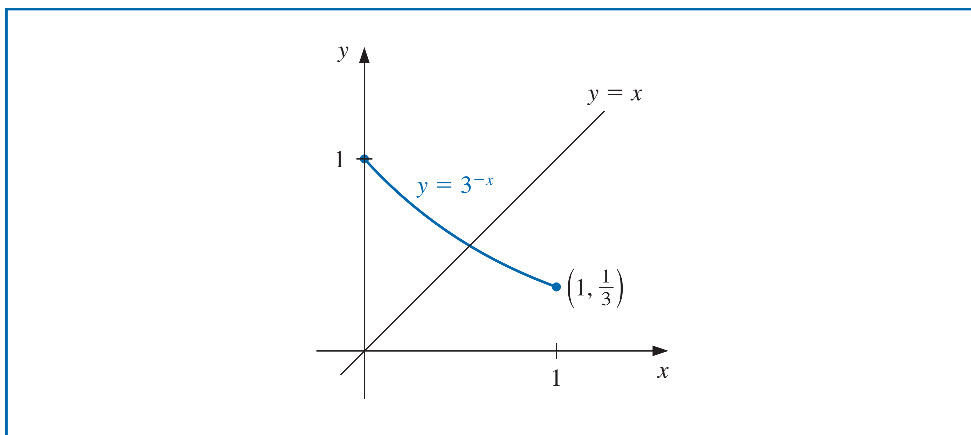
Por lo tanto, para $x \in [0, 1]$, tenemos $g(x) \in [0, 1]$. La primera parte del teorema 2.3 garantiza que existe por lo menos un punto fijo en $[0, 1]$.

Sin embargo,

$$g'(0) = -\ln 3 = -1.098612289,$$

por lo que $|g'(x)| \not\leq 1$ en $(0, 1)$, y el teorema 2.3 no se puede usar para determinar la unicidad. Pero g siempre decrece y es claro, a partir de la figura 2.5, que el punto fijo debe ser único. ■

Figura 2.5



Iteración de punto fijo

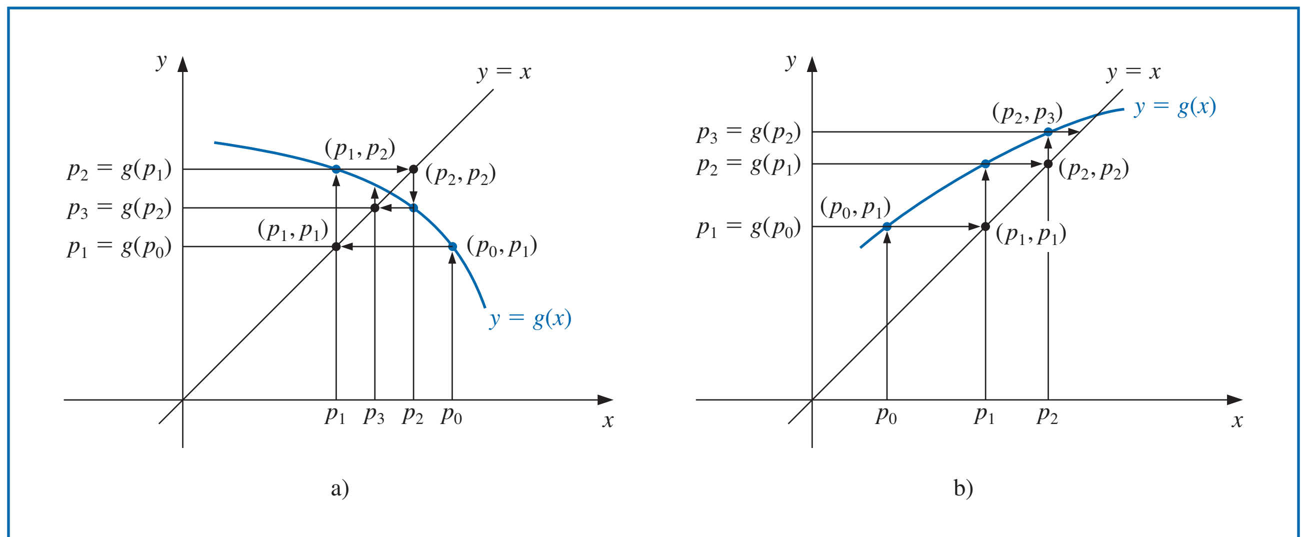
No podemos determinar explícitamente el punto fijo en el ejemplo 3 porque no tenemos otra forma de resolver la ecuación $p = g(p) = 3^{-p}$. Sin embargo, podemos determinar las aproximaciones para este punto fijo con cualquier grado específico de precisión. Ahora consideraremos la forma de hacerlo.

Para aproximar el punto fijo de una función g, elegimos una aproximación inicial p_0 y generamos la sucesión $\{p_n\}_{n=0}^{\infty}$ al permitir $p_n = g(p_{n-1})$, para cada $n \geq 1$. Si la sucesión converge a p y g es continua, entonces

$$p = \lim_{n\to\infty} p_n = \lim_{n\to\infty} g(p_{n-1}) = g\left(\lim_{n\to\infty} p_{n-1}\right) = g(p),$$

y se obtiene una solución para $x = g(x)$. Esta técnica recibe el nombre de **punto fijo** o **iteración funcional**. El procedimiento se ilustra en la figura 2.6 y se detalla en el algoritmo 2.2.

Figura 2.6



ALGORITMO **2.2**

Iteración de punto fijo

Encontrar una solución de $p = g(p)$ dada una aproximación inicial p_0:

ENTRADA aproximación inicial p_0; tolerancia TOL; número máximo de iteraciones N_0 .

SALIDA aproximada p o mensaje de falla.

Paso 1 Determine $i = 1$.

Paso 2 Mientras $i \leq N_0$ haga los pasos 3–6.

Paso 3 Determine $p = g(p_0)$. (*Calcule* p_i.)

Paso 4 Si $|p - p_0| < TOL$ entonces
SALIDA (p); (*El procedimiento fue exitoso*.)
PARE.

Paso 5 Determine $i = i + 1$.

Paso 6 Determine $p_0 = p$. (*Actualizar* p_0.)

Paso 7 SALIDA ('El método falló después de N_0 iteraciones, N_0 =', N_0);
(*El procedimiento no fue exitoso*.)
PARE. ■

Lo siguiente ilustra algunas características de la iteración funcional.

Ilustración La ecuación $x^3 + 4x^2 - 10 = 0$ tiene una raíz única en [1, 2]. Existen muchas formas de cambiar la ecuación para la forma de punto fijo $x = g(x)$ mediante una simple manipulación algebraica. Por ejemplo, para obtener la función g descrita en la parte c), podemos manipular la ecuación $x^3 + 4x^2 - 10 = 0$ como sigue:

$$4x^2 = 10 - x^3, \quad \text{por lo que} \quad x^2 = \frac{1}{4}(10 - x^3), \quad \text{y} \quad x = \pm\frac{1}{2}(10 - x^3)^{1/2}.$$

Para obtener una solución positiva, se selecciona $g_3(x)$. No es importante derivar las funciones que se muestran aquí, pero debemos verificar que el punto fijo de cada una es en realidad una solución para la ecuación original, $x^3 + 4x^2 - 10 = 0$.

a) $x = g_1(x) = x - x^3 - 4x^2 + 10$

b) $x = g_2(x) = \left(\frac{10}{x} - 4x\right)^{1/2}$

c) $x = g_3(x) = \frac{1}{2}(10 - x^3)^{1/2}$

d) $x = g_4(x) = \left(\frac{10}{4 + x}\right)^{1/2}$

e) $x = g_5(x) = x - \frac{x^3 + 4x^2 - 10}{3x^2 + 8x}$

Con $p_0 = 1.5$, la tabla 2.2 enumera los resultados de la iteración de punto fijo para las cinco selecciones de g.

Tabla 2.2

n	(a)	(b)	(c)	(d)	(e)
0	1.5	1.5	1.5	1.5	1.5
1	−0.875	0.8165	1.286953768	1.348399725	1.373333333
2	6.732	2.9969	1.402540804	1.367376372	1.365262015
3	−469.7	$(-8.65)^{1/2}$	1.345458374	1.364957015	1.365230014
4	1.03×10^8		1.375170253	1.365264748	1.365230013
5			1.360094193	1.365225594	
6			1.367846968	1.365230576	
7			1.363887004	1.365229942	
8			1.365916734	1.365230022	
9			1.364878217	1.365230012	
10			1.365410062	1.365230014	
15			1.365223680	1.365230013	
20			1.365230236		
25			1.365230006		
30			1.365230013		

La raíz real es 1.365230013, como se observó en el ejemplo 1 de la sección 2.1. Al comparar los resultados con el algoritmo de bisección provisto en ese ejemplo, se puede observar que se han obtenido excelentes resultados para las opciones d) y e) (el método de bisección requiere 27 iteraciones para esta precisión). Es interesante observar que la opción a) era divergente y que la opción b) se volvió indefinida porque involucra la raíz cuadrada de un número negativo. ■

Aunque las diferentes funciones que hemos proporcionado son problemas de punto fijo para el mismo problema de encontrar la raíz, varían ampliamente como técnicas para aproximar la solución a este último. Su objetivo es ilustrar lo que se debe contestar:

- Pregunta: ¿Cómo podemos encontrar un problema de punto fijo que produce una sucesión en la que la fiabilidad y la rapidez convergen a la solución del problema de encontrar la raíz determinada?

El siguiente teorema y su corolario nos proporcionan algunas claves respecto a los procedimientos que deberíamos seguir y, tal vez más importante, algunos que deberíamos rechazar.

Teorema 2.4 **(Teorema de punto fijo)**

Sea $g \in C[a, b]$ tal que $g(x) \in [a, b]$ para todas las x en $[a, b]$. Suponga, además, que existe g' en (a, b) y que existe una constante $0 < k < 1$ con

$$|g'(x)| \leq k, \quad \text{para todas} \quad x \in (a, b).$$

Entonces, para cualquier número p_0 en [a, b], la sucesión definida por

$$p_n = g(p_{n-1}), \quad n \geq 1,$$

converge al único punto fijo p en $[a, b]$.

Demostración El teorema 2.3 implica que existe un único punto p en $[a, b]$ con $g(p) = p$. Ya que g mapea $[a, b]$ en sí mismo, la sucesión $\{p_n\}_{n=0}^{\infty}$ se define para todas las $n \geq 0$, y $p_n \in [a, b]$ para todas las n. Al utilizar el hecho de que $|g'(x)| \leq k$ y el teorema de valor medio 1.8, tenemos, para cada n,

$$|p_n - p| = |g(p_{n-1}) - g(p)| = |g'(\xi_n)||p_{n-1} - p| \leq k|p_{n-1} - p|,$$

donde $\xi_n \in (a, b)$. Al aplicar esta desigualdad de manera inductiva obtenemos

$$|p_n - p| \leq k|p_{n-1} - p| \leq k^2|p_{n-2} - p| \leq \cdots \leq k^n|p_0 - p|. \tag{2.4}$$

Ya que $0 < k < 1$, tenemos $\text{lím}_{n\to\infty} k^n = 0$ y

$$\lim_{n\to\infty} |p_n - p| \leq \lim_{n\to\infty} k^n|p_0 - p| = 0.$$

Por lo tanto, $\{p_n\}_{n=0}^{\infty}$ converge a p. ■

Corolario 2.5 Si g satisface las hipótesis del teorema 2.4, entonces las cotas del error relacionado con el uso de p_n para aproximar p, están dadas por por

$$|p_n - p| \leq k^n \text{ máx}\{p_0 - a, b - p_0\} \tag{2.5}$$

y

$$|p_n - p| \leq \frac{k^n}{1-k}|p_1 - p_0|, \quad \text{para toda } n \geq 1. \tag{2.6}$$

Demostración Puesto que $p \in [a, b]$ la primera cota se sigue la desigualdad (2.4):

$$|p_n - p| \leq k^n|p_0 - p| \leq k^n \text{ máx}\{p_0 - a, b - p_0\}.$$

Para $n \geq 1$, el procedimiento que se usa en la prueba del teorema 2.4 implica que

$$|p_{n+1} - p_n| = |g(p_n) - g(p_{n-1})| \leq k|p_n - p_{n-1}| \leq \cdots \leq k^n|p_1 - p_0|.$$

Por lo tanto, para $m > n \geq 1$,

$$\begin{aligned}|p_m - p_n| &= |p_m - p_{m-1} + p_{m-1} - \cdots + p_{n+1} - p_n| \\ &\leq |p_m - p_{m-1}| + |p_{m-1} - p_{m-2}| + \cdots + |p_{n+1} - p_n| \\ &\leq k^{m-1}|p_1 - p_0| + k^{m-2}|p_1 - p_0| + \cdots + k^n|p_1 - p_0| \\ &= k^n|p_1 - p_0|\left(1 + k + k^2 + \cdots + k^{m-n-1}\right).\end{aligned}$$

Mediante el teorema 2.3, $\lim_{m\to\infty} p_m = p$, por lo que

$$|p - p_n| = \lim_{m\to\infty} |p_m - p_n| \leq \lim_{m\to\infty} k^n|p_1 - p_0| \sum_{i=0}^{m-n-1} k^i \leq k^n|p_1 - p_0| \sum_{i=0}^{\infty} k^i.$$

Pero $\sum_{i=0}^{\infty} k^i$ es una serie geométrica con radio k y $0 < k < 1$. Esta sucesión converge a $1/(1 - k)$, lo que nos da la segunda cota:

$$|p - p_n| \leq \frac{k^n}{1 - k}|p_1 - p_0|. \qquad ■$$

Ambas desigualdades en el corolario relacionan la razón en la que $\{p_n\}_{n=0}^{\infty}$ converge a la cota k de la primera derivada. La razón de convergencia depende del factor k^n. Mientras más pequeño sea el valor de k, más rápido convergerá. Sin embargo, la convergencia sería muy lenta si k está cercana a 1.

Ilustración Reconsideremos los diferentes esquemas de punto fijo descritos en la ilustración anterior en vista del teorema de punto fijo 2.4 y su corolario 2.5.

a) Para $g_1(x) = x - x^3 - 4x^2 + 10$, tenemos $g_1(1) = 6$ y $g_1(2) = -12$, por lo que g_1 no mapea $[1, 2]$ en sí mismo. Además $g'_1(x) = 1 - 3x^2 - 8x$, por lo que $|g'_1(x)| > 1$ para todas las x en $[1, 2]$. A pesar de que el teorema 2.4 no garantiza que el método debe fallar para esta selección de g, no existe una razón para esperar convergencia.

b) Con $g_2(x) = [(10/x) - 4x]^{1/2}$, podemos ver que g_2 no traza un mapa $[1, 2]$ en $[1, 2]$, y la sucesión $\{p_n\}_{n=0}^{\infty}$ no está definida cuando $p_0 = 1.5$. Además, no existe un intervalo que contiene $p \approx 1.365$ tal que $|g'_2(x)| < 1$ puesto que $|g'_2(p)| \approx 3.4$. No existe una razón para esperar que este método convergerá.

c) Para la función $g_3(x) = \frac{1}{2}(10 - x^3)^{1/2}$ tenemos

$$g'_3(x) = -\frac{3}{4}x^2(10 - x^3)^{-1/2} < 0 \quad \text{en } [1, 2],$$

por lo que g_3 es estrictamente decreciente en $[1, 2]$. Sin embargo $|g'_3(2)| \approx 2.12$, por lo que la condición $|g'_3(x)| \leq k < 1$ falla en $[1, 2]$. Un análisis más cercano de la sucesión $\{p_n\}_{n=0}^{\infty}$ comenzando con $p_0 = 1.5$ muestra que es suficiente considerar el intervalo $[1, 1.5]$ en lugar de $[1, 2]$. En este intervalo, sigue siendo verdad que $g'_3(x) < 0$ y g_3 es estrictamente decreciente, pero, además,

$$1 < 1.28 \approx g_3(1.5) \leq g_3(x) \leq g_3(1) = 1.5,$$

para todas las $x \in [1, 1.5]$. Esto muestra que g_3 mapea el intervalo $[1, 1.5]$ en sí mismo. También es verdad que $|g'_3(x)| \leq |g'_3(1.5)| \approx 0.66$ en este intervalo, por lo que el teorema 2.4 confirma la convergencia que ya conocemos.

d) Para $g_4(x) = (10/(4+x))^{1/2}$, tenemos

$$|g_4'(x)| = \left|\frac{-5}{\sqrt{10}(4+x)^{3/2}}\right| \leq \frac{5}{\sqrt{10}(5)^{3/2}} < 0.15, \quad \text{para todas las} \quad x \in [1, 2].$$

La magnitud de la cota de $g'_4(x)$ es mucho menor que la magnitud de la cota (encontrada en c)) $g'_3(x)$, lo cual explica la convergencia más rápida usando g_4.

e) La sucesión definida por

$$g_5(x) = x - \frac{x^3 + 4x^2 - 10}{3x^2 + 8x}$$

converge mucho más rápido que nuestras otras elecciones. En las siguientes secciones, observaremos de dónde proviene esta elección y porqué es tan efectiva. ■

A partir de lo que hemos observado, la

- Pregunta: ¿Cómo podemos encontrar un problema de punto fijo que produce una sucesión que converge de manera confiable y rápida en una solución para un problema dado de encontrar la raíz?

podríamos tener la

- Respuesta: manipular el problema de encontrar la raíz en un problema de punto fijo que satisfaga las condiciones del teorema de punto fijo 2.4 y hacer la derivada tan pequeña como sea posible cerca del punto fijo.

En las siguientes secciones examinaremos esto con mayor detalle.

La sección Conjunto de ejercicios 2.2 está disponible en línea. Encuentre la ruta de acceso en las páginas preliminares.



2.3 Método de Newton y sus extensiones

Isaac Newton (1641–1727) fue uno de los científicos más brillantes de todos los tiempos. El final del siglo XVII fue un periodo vibrante para la ciencia y las matemáticas, y el trabajo de Newton tocó casi todos los aspectos de esta última ciencia. Se presentó su método de resolución para encontrar la raíz de la ecuación $y^3 - 2y - 5 = 0$. A pesar de que demostró el método sólo para polinomios, es claro que conocía sus aplicaciones más amplias.

El **método de Newton** (o de *Newton-Raphson*) es uno de los métodos numéricos más poderosos y reconocidos para resolver un problema de encontrar la raíz. Existen muchas formas de presentar el método de Newton.

Método de Newton

Si sólo queremos un algoritmo, podemos considerar la técnica de manera gráfica, como a menudo se hace en cálculo. Otra posibilidad es derivar el método de Newton como una técnica para obtener convergencia más rápida de lo que ofrecen otros tipos de iteración funcional, como hacemos en la sección 2.4. Una tercera forma para presentar el método de Newton, que se analiza a continuación, está basada en los polinomios de Taylor. Ahí observaremos que esta forma particular no sólo produce el método, sino también una cota para el error de aproximación.

Suponga que $f \in C^2[a, b]$. Si $p_0 \in [a, b]$ es una aproximación para p, de tal forma que $f'(p_0) \neq 0$ y $|p - p_0|$ es "pequeño". Considere que el primer polinomio de Taylor para $f(x)$ expandido alrededor de p_0 y evaluado en $x = p$:

$$f(p) = f(p_0) + (p - p_0)f'(p_0) + \frac{(p - p_0)^2}{2} f''(\xi(p)),$$

donde $\xi(p)$ se encuentra entre p y p_0. Puesto que $f(p) = 0$, esta ecuación nos da

$$0 = f(p_0) + (p - p_0)f'(p_0) + \frac{(p - p_0)^2}{2} f''(\xi(p)).$$

El método de Newton se deriva al suponer que como $|p - p_0|$ es pequeño, el término relacionado con $(p - p_0)^2$ es mucho más pequeño, entonces

$$0 \approx f(p_0) + (p - p_0)f'(p_0).$$

Joseph Raphson (1648–1715) proporcionó una descripción del método atribuido a Isaac Newton en 1690 y reconoció a Newton como la fuente del descubrimiento. Ni Newton ni Raphson utilizaron explícitamente la derivada en su descripción ya que ambos sólo consideraron polinomios. Otros matemáticos, en especial James Gregory (1636–1675), estaban conscientes del proceso subyacente en esa época o antes.

Al resolver para p obtenemos

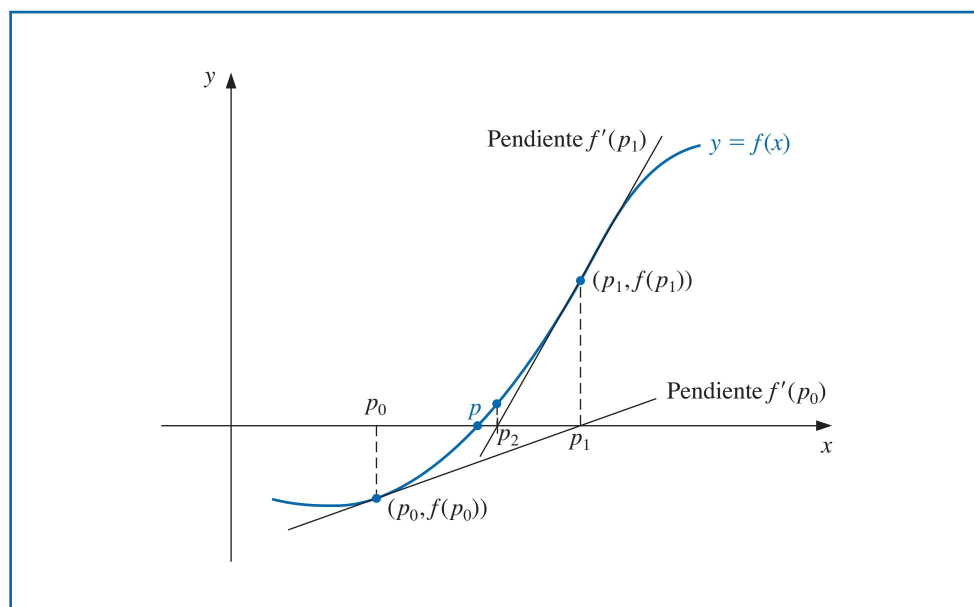
$$p \approx p_0 - \frac{f(p_0)}{f'(p_0)} \equiv p_1.$$

Esto constituye la base para el método de Newton, que empieza con una aproximación inicial p_0 y genera la sucesión $\{p_n\}_{n=0}^{\infty}$, mediante

$$p_n = p_{n-1} - \frac{f(p_{n-1})}{f'(p_{n-1})}, \quad \text{para } n \geq 1. \tag{2.7}$$

La figura 2.7 ilustra cómo se obtienen las aproximaciones usando tangentes sucesivas. (Además observe el ejercicio 31.) Al empezar con la aproximación inicial p_0, la aproximación p_1 es la intersección con el eje x de la recta tangente a la gráfica de f en $(p_0, f(p_0))$. La aproximación p_2 es la intersección con el eje x de la recta tangente a la gráfica f en $(p_1, f(p_1))$ y así sucesivamente. El algoritmo 2.3 implementa este procedimiento.

Figura 2.7



ALGORITMO **2.3**

Método de Newton

Para encontrar una solución a $f(x) = 0$ dada una aproximación inicial p_0:

ENTRADA aproximación inicial p_0 tolerancia TOL; número máximo de iteraciones N_0

SALIDA solución aproximada p o mensaje de falla.

Paso 1 Determine $i = 1$.



Paso 2 Mientras $i \leq N_0$ haga los pasos 3–6.

Paso 3 Determine $p = p_0 - f(p_0)/f'(p_0)$. (*Calcule* p_i.)

Paso 4 Si $|p - p_0| < TOL$ entonces
SALIDA (p); (*El procedimiento fue exitoso*.)
PARE.

Paso 5 Determine $i = i + 1$.

Paso 6 Determine $p_0 = p$. (*Actualizce* p_0.)

Paso 7 SALIDA ('El método falló después de N_0 iteraciones, $N_0 =$', N_0);
(*El procedimiento no fue exitoso*.)
PARE. ■

Las desigualdades de la técnica de parada determinadas con el método de bisección son aplicables al método de Newton. Es decir, seleccione una tolerancia $\varepsilon > 0$ y construya p_1, …, p_N hasta que

$$|p_N - p_{N-1}| < \varepsilon, \tag{2.8}$$

$$\frac{|p_N - p_{N-1}|}{|p_N|} < \varepsilon, \quad p_N \neq 0, \tag{2.9}$$

o

$$|f(p_N)| < \varepsilon. \tag{2.10}$$

Una forma de la desigualdad (2.8) se usa en el paso 4 del algoritmo 2.3. Observe que ninguna de las desigualdades (2.8), (2.9) o (2.10) dan información precisa sobre el error real $|p_N - p|$. (Consulte los ejercicios 19 y 20 en la sección 2.1).

El método de Newton es una técnica de iteración funcional con $p_n = g(p_{n-1})$, para la que

$$g(p_{n-1}) = p_{n-1} - \frac{f(p_{n-1})}{f'(p_{n-1})}, \quad \text{para } n \geq 1. \tag{2.11}$$

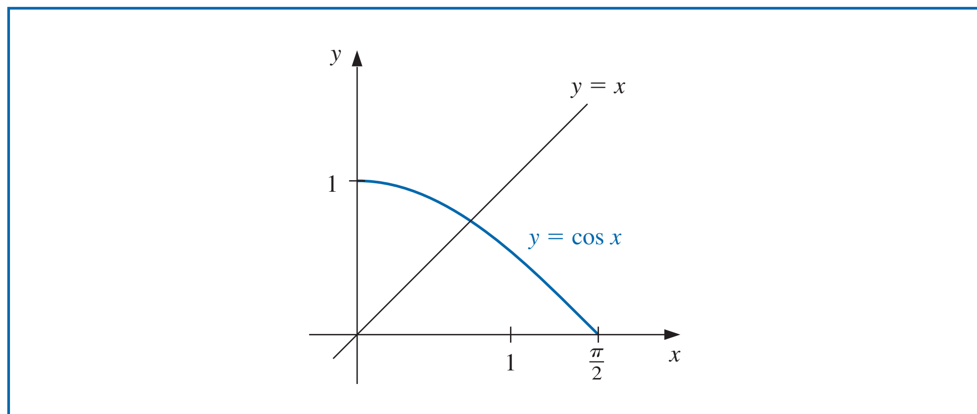
De hecho, esta es la técnica de iteración funcional que se usó para proporcionar la convergencia rápida que vimos en la columna e) de la tabla 2.2 en la sección 2.2.

A partir de la ecuación (2.7) es claro que el método de Newton no se puede continuar si $f'(p_{n-1}) = 0$ para alguna n. De hecho, observaremos que el método es más eficaz cuando f' está acotada lejos de cero y cerca de p.

Ejemplo 1 Considere la función $f(x) = \cos x - x = 0$. Aproxime una raíz de f usando **a)** el método de punto fijo y **b)** el método de Newton.

Solución **a)** Una solución para este problema de encontrar la raíz también es una solución para el problema de punto fijo $x = \cos x$, y la gráfica en la figura 2.8 implica que un solo punto fijo p se encuentra en $[0, \pi/2]$.

Figura 2.8



Observe que la variable en la función trigonométrica está medida en radianes, no grados. Éste siempre será el caso a menos que se especifique lo contrario.

Tabla 2.3

n	p_n
0	0.7853981635
1	0.7071067810
2	0.7602445972
3	0.7246674808
4	0.7487198858
5	0.7325608446
6	0.7434642113
7	0.7361282565

La tabla 2.3 muestra los resultados de la iteración de punto fijo con $p_0 = \pi/4$. Lo mejor que podríamos concluir a partir de estos resultados es que $p \approx 0.74$.

b) Para aplicar el método de Newton a este problema, necesitamos $f'(x) = -\text{sen}\, x - 1$. Al iniciar de nuevo en $p_0 = \pi/4$, tenemos

$$
\begin{aligned}
p_1 &= p_0 - \frac{p_0}{f'(p_0)} \\
&= \frac{\pi}{4} - \frac{\cos(\pi/4) - \pi/4}{-\text{sen}(\pi/4) - 1} \\
&= \frac{\pi}{4} - \frac{\sqrt{2}/2 - \pi/4}{-\sqrt{2}/2 - 1} \\
&= 0.7395361337 \\
p_2 &= p_1 - \frac{\cos(p_1) - p_1}{-\text{sen}(p_1) - 1} \\
&= 0.7390851781
\end{aligned}
$$

Nosotros generamos continuamente la sucesión mediante

$$p_n = p_{n-1} - \frac{f(p_{n-1})}{f'(p_{n-1})} = p_{n-1} - \frac{\cos p_{n-1} - p_{n-1}}{-\text{sen}\, p_{n-1} - 1}.$$

Tabla 2.4

Método de Newton

n	p_n
0	0.7853981635
1	0.7395361337
2	0.7390851781
3	0.7390851332
4	0.7390851332

Esto nos da las aproximaciones en la tabla 2.4. Se obtiene una excelente aproximación con $n = 3$. Debido a la cercanía de p_3 y p_4, podríamos esperar razonablemente que este resultado sea preciso con los decimales enumerados. ■

Convergencia con el método de Newton

El ejemplo 1 muestra que el método de Newton puede dar aproximaciones en extremo precisas con muy pocas iteraciones. Para ese ejemplo, sólo se necesitó una iteración del método de Newton para dar una precisión mayor que las siete iteraciones del método de punto fijo. Ahora es momento de examinar el método de Newton para descubrir porqué es tan eficaz.

La derivación del método de Newton por medio de la serie de Taylor al inicio de la sección señala la importancia de una aproximación inicial precisa. La suposición crucial es que el término relacionado con $(p - p_0)^2$ es, en comparación con $|p - p_0|$, tan pequeño que se puede eliminar. Claramente esto será falso a menos que p_0 sea una buena aproximación para p. Si p_0 no está suficientemente cerca de la raíz real, existen pocas razones para sospechar que el método de Newton convergerá en la raíz. Sin embargo, en algunos casos, incluso las malas aproximaciones iniciales producirán convergencia. (Los ejercicios 15 y 16 ilustran algunas de las posibilidades.)

El siguiente teorema de convergencia para el método de Newton ilustra la importancia teórica de la selección de p_0.

Teorema 2.6 Sea $f \in C^2[a, b]$. Si $p \in (a, b)$ es tal que $f(p) = 0$ y $f'(p) \neq 0$, entonces existe una $\delta > 0$ tal que el método de Newton genera una sucesión $\{p_n\}_{n=1}^{\infty}$ que converge a p para cualquier aproximación inicial $p_0 \in [p - \delta, p + \delta]$.

Demostración La prueba se basa en el análisis del método de Newton como un esquema de iteración funcional $p_n = g(p_{n-1})$ para $n \geq 1$, con

$$g(x) = x - \frac{f(x)}{f'(x)}.$$

Sea k un número entre $(0, 1)$. Primero encontramos un intervalo $[p - \delta, p + \delta]$ que g mapea en sí mismo y para el cual $|g'(x)| \leq k$ para todas las $x \in (p - \delta, p + \delta)$.

Ya que f' es continua y $f'(p) \neq 0$, la parte a) del ejercicio 30 en la sección 1.1 implica que existe una $\delta_1 > 0$, tal que $f'(x) \neq 0$ para $x \in [p - \delta_1, p + \delta_1] \subseteq [a, b]$. Por lo tanto, g está definida y es continua en $[p - \delta_1, p + \delta_1]$. Además,

$$g'(x) = 1 - \frac{f'(x)f'(x) - f(x)f''(x)}{[f'(x)]^2} = \frac{f(x)f''(x)}{[f'(x)]^2},$$

para $x \in [p - \delta_1, p + \delta_1]$ y, puesto que $f \in C^2[a, b]$, tenemos $g \in C^1[p - \delta_1, p + \delta_1]$.

Por hipótesis, $f(p) = 0$ por lo que

$$g'(p) = \frac{f(p)f''(p)}{[f'(p)]^2} = 0.$$

Puesto que g' es continua y $0 < k < 1$, la parte b) del ejercicio 30 en la sección 1.1 implica que existe δ, con $0 < \delta < \delta_1$, para el cual

$$|g'(x)| \leq k, \qquad \text{para todas las } x \in [p - \delta, p + \delta].$$

Falta probar que g mapea $[p - \delta, p + \delta]$ en $[p - \delta, p + \delta]$. Si $x \in [p - \delta, p + \delta]$, el teorema de valor medio implica que para algún número ξ entre x y p, $|g(x) - g(p)| = |g'(\xi)||x - p|$. Por lo tanto,

$$|g(x) - p| = |g(x) - g(p)| = |g'(\xi)||x - p| \leq k|x - p| < |x - p|.$$

Puesto que $x \in [p - \delta, p + \delta]$, se sigue que $|x - p| < \delta$ y que $|g(x) - p| < \delta$. Por lo tanto, g mapea $[p - \delta, p + \delta]$ en $[p - \delta, p + \delta]$.

Ahora, se satisfacen todas las hipótesis del teorema de punto fijo 2.4, por lo que la sucesión $\{p_n\}_{n=1}^{\infty}$, definida por

$$p_n = g(p_{n-1}) = p_{n-1} - \frac{f(p_{n-1})}{f'(p_{n-1})}, \qquad \text{para } n \geq 1,$$

converge a p para cualquier $p_0 \in [p - \delta, p + \delta]$. ■

El teorema 2.6 establece que, de acuerdo con suposiciones razonables, el método de Newton converge, siempre que se seleccione una aproximación inicial suficientemente exacta. También implica que la constante k que acota la derivada de g y, por consiguiente, indica que la velocidad de convergencia del método disminuye a 0, conforme el procedimiento continúa. Este resultado es importante para la teoría del método de Newton, pero casi nunca se aplica en la práctica porque no nos dice cómo determinar δ.

En una aplicación práctica, se selecciona una aproximación inicial y se generan aproximaciones sucesivas con el método de Newton. Ya sea que éstos converjan rápidamente a la raíz o será claro que la convergencia es poco probable.

El método de la secante

El método de Newton es una técnica en extremo poderosa, pero tiene una debilidad importante: la necesidad de conocer el valor de la derivada de f en cada aproximación. Con frecuencia, $f'(x)$ es mucho más difícil y necesita más operaciones aritméticas para calcular $f(x)$.

Para evitar el problema de la evaluación de la derivada en el método de Newton, presentamos una ligera variación. Por definición,

$$f'(p_{n-1}) = \lim_{x \to p_{n-1}} \frac{f(x) - f(p_{n-1})}{x - p_{n-1}}.$$

Si p_{n-2} está cerca de p_{n-1}, entonces

$$f'(p_{n-1}) \approx \frac{f(p_{n-2}) - f(p_{n-1})}{p_{n-2} - p_{n-1}} = \frac{f(p_{n-1}) - f(p_{n-2})}{p_{n-1} - p_{n-2}}.$$

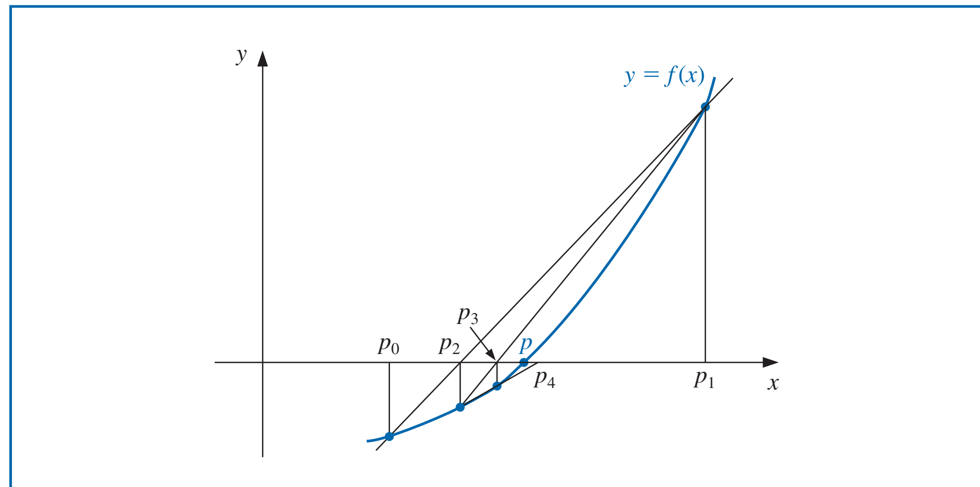
Usando esta aproximación para $f'(p_{n-1})$ en la fórmula de Newton obtenemos

La palabra *secante* se deriva de la palabra en latín *secan*, que significa "cortar". El método de la secante usa una línea secante, que une dos puntos que cortan la curva, para aproximar una raíz.

$$p_n = p_{n-1} - \frac{f(p_{n-1})(p_{n-1} - p_{n-2})}{f(p_{n-1}) - f(p_{n-2})}. \tag{2.12}$$

Esta técnica recibe el nombre de método de la secante y se presenta en el algoritmo 2.4 (consulte la figura 2.9). Empezando con dos aproximaciones iniciales p_0 y p_1, la aproximación p_2 es la intersección en x de la recta que une los puntos $(p_0, f(p_0))$ y $(p_1, f(p_1))$. La aproximación p_3 es la intersección en x de la recta que une los puntos $(p_1, f(p_1))$ y $(p_2, f(p_2))$ y así sucesivamente. Observe que sólo se necesita una evaluación de la función por cada paso para el método de la secante después de haber determinado p_2. En contraste, cada paso del método de Newton requiere una evaluación tanto de la función como de su derivada.

Figura 2.9



ALGORITMO 2.4

Método de la secante

Para encontrar una solución para $f(x) = 0$ dadas las aproximaciones iniciales p_0 y p_1:

ENTRADA aproximaciones iniciales p_0, p_1 tolerancia TOL; número máximo de iteraciones N_0.

SALIDA solución aproximada p o mensaje de falla.

Paso 1 Determine $i = 2$;
$q_0 = f(p_0)$;
$q_1 = f(p_1)$.



Paso 2 Mientras $i \leq N_0$ haga los pasos 3–6.

Paso 3 Determine $p = p_1 - q_1(p_1 - p_0)/(q_1 - q_0)$. (*Calcule* p_i.)

Paso 4 Si $|p - p_1| < TOL$ entonces
SALIDA (p); (*El procedimiento fue exitoso*.)
PARE.

Paso 5 Determine $i = i + 1$.

Paso 6 Determine $p_0 = p_1$; (*Actualice* p_0, q_0, p_1, q_1.)
$q_0 = q_1$;
$p_1 = p$;
$q_1 = f(p)$.

Paso 7 SALIDA ('El método falló después de N_0 iteraciones, N_0 =', N_0);
(*El procedimiento no fue exitoso*.)
PARE. ■

El siguiente ejemplo incluye el problema que se consideró en el ejemplo 1, donde utilizamos el método de Newton con $p_0 = \pi/4$.

Ejemplo 2 Use el método de la secante para encontrar una solución para $x = \cos x$ y compare las aproximaciones con las determinadas en el ejemplo 1, el cual aplica el método de Newton.

Solución En el ejemplo 1, comparamos la iteración del punto fijo y el método de Newton con la aproximación inicial $p_0 = \pi/4$. Para el método de la secante, necesitamos dos aproximaciones iniciales.

Suponga que usamos $p_0 = 0.5$ y $p_1 = \pi/4$:

$$p_2 = p_1 - \frac{(p_1 - p_0)(\cos p_1 - p_1)}{(\cos p_1 - p_1) - (\cos p_2 - p_2)}$$
$$= \frac{\pi}{4} - \frac{(\pi/4 - 0.5)(\cos(\pi/4) - \pi/4)}{(\cos(\pi/4) - \pi/4) - (\cos 0.5 - 0.5)}$$
$$= 0.7363841388.$$

Las aproximaciones sucesivas se generan con la fórmula

$$p_n = p_{n-1} - \frac{(p_{n-1} - p_{n-2})(\cos p_{n-1} - p_{n-1})}{(\cos p_{n-1} - p_{n-1}) - (\cos p_{n-2} - p_{n-2})}, \quad \text{para } n \geq 2.$$

Esto proporciona los resultados en la tabla 2.5. Observamos que a pesar de que la fórmula para p_2 parece indicar un cálculo repetido, una vez que se calcula $f(p_0)$ y $f(p_1)$, no se calculan de nuevo. ■

Tabla 2.5

	Secante
n	p_n
0	0.5
1	0.7853981635
2	0.7363841388
3	0.7390581392
4	0.7390851493
5	0.7390851332

	Newton
n	p_n
0	0.7853981635
1	0.7395361337
2	0.7390851781
3	0.7390851332
4	0.7390851332

Al comparar los resultados en la tabla 2.5 a partir del método de la secante y el método de Newton, observamos que la aproximación p_5 del método de la secante es precisa hasta la décima cifra decimal, mientras que con el método de Newton se obtuvo esta precisión por p_3. Para este ejemplo, la convergencia del método de la secante es mucho más rápida que la iteración funcional, pero ligeramente más lenta que el método de Newton. Éste es, en general, el caso. (Consulte el ejercicio 14 de la sección 2.4.)

El método de Newton o el método de la secante con frecuencia se usan para refinar una respuesta obtenida con otra técnica, como el método de bisección, ya que estos métodos requieren una primera aproximación adecuada pero, en general, proveen convergencia rápida.

El método de posición falsa

Cada par sucesivo de aproximaciones en el método de bisección agrupa una raíz p de la ecuación; es decir, para cada entero positivo n, se encuentra una raíz entre a_n y b_n. Esto implica que, para cada n, las iteraciones del método de bisección satisfacen

$$|p_n - p| < \frac{1}{2}|a_n - b_n|,$$

lo cual provee una cota del error fácilmente calculable para las aproximaciones.

La agrupación de raíces no está garantizada para el método de Newton ni para el método de la secante. En el ejemplo 1, el método de Newton se aplicó a $f(x) = \cos x - x$ y se encontró que la raíz aproximada es 0.7390851332. La tabla 25 muestra que esta raíz no se agrupa mediante p_0 *y* p_1 *o* p_1 *y* p_2. Las aproximaciones del método de la secante para este problema también se determinan en la tabla 2.5. En este caso, las aproximaciones iniciales p_0 *y* p_1 agrupan la raíz, pero el par de aproximaciones p_3 *y* p_4 fallan al hacerlo.

El término *Regula Falsi*, literalmente "regla falsa" o "posición falsa" se refiere a una técnica en la que se usan resultados que se sabe son falsos, pero de algún modo específico, para obtener convergencia a un resultado verdadero. Los problemas de posición falsa se pueden encontrar en el papiro Rhind, que data de aproximadamente 1650 a.C.

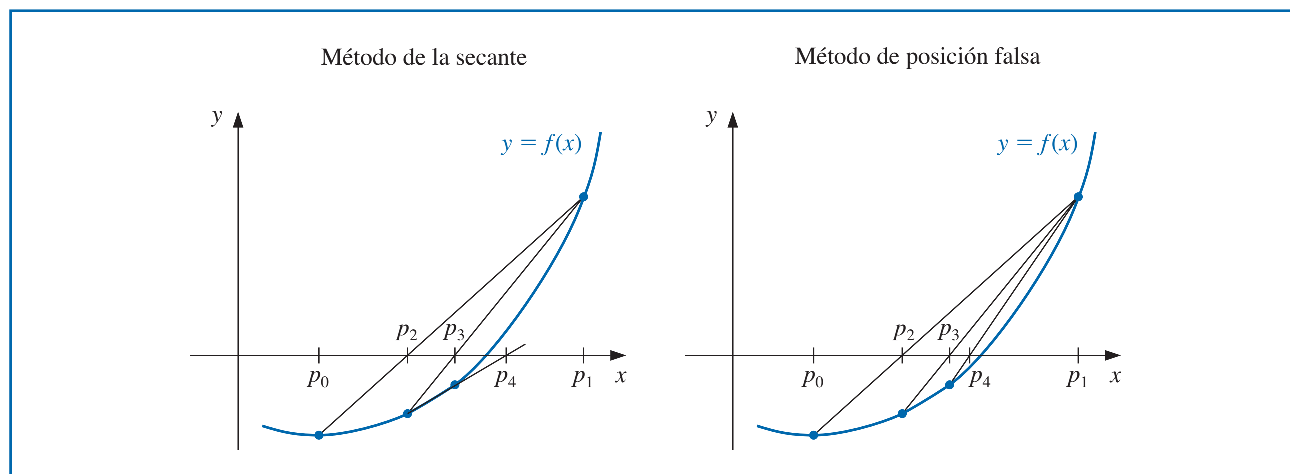
El **método de posición falsa** (también llamado *Regula Falsi*) genera aproximaciones de la misma manera que el método de la secante, pero incluye una prueba para garantizar que la raíz siempre se agrupa entre iteraciones sucesivas. A pesar de que no es un método que por lo general recomendamos, ilustra cómo se puede integrar la agrupación.

En primer lugar, seleccionamos las aproximaciones iniciales p_0 y p_1 con $f(p_0) \cdot f(p_1) < 0$. La aproximación p_2 se selecciona de la misma forma que en el método de la secante como la intersección en x de la recta que une $(p_0, f(p_0))$ y $(p_1, f(p_1))$. Para decidir cuál línea secante se usa para calcular p_3, considere $f(p_2) \cdot f(p_1)$ o, más correctamente, $\operatorname{sgn} f(p_2) \cdot \operatorname{sgn} f(p_1)$.

- Si $\operatorname{sgn} f(p_2) \cdot \operatorname{sgn} f(p_1) < 0$, entonces p_1 y p_2 agrupan una raíz. Seleccione p_3 como la intersección en x de la recta que une $(p_1, f(p_1))$ y $(p_2, f(p_2))$.
- Si no, seleccionamos p_3 como la intersección en x de la recta que une $(p_0, f(p_0))$ y $(p_2, f(p_2))$ y, a continuación intercambia los índices en p_0 y p_1.

De manera similar, una vez que se encuentra p_3, el signo de $f(p_3) \cdot f(p_2)$ determina si usamos p_2 y p_3 o p_3 y p_1 para calcular p_4. En el último caso, se vuelve a etiquetar p_2 y p_1. Reetiquetar garantiza que la raíz se agrupa entre iteraciones sucesivas. El proceso, como se describe en el algoritmo 2.5 y la figura 2.10 muestra cómo las iteraciones pueden diferir de las del método de la secante. En esta ilustración, las primeras tres aproximaciones son iguales, pero la cuarta difiere.

Figura 2.10



ALGORITMO 2.5

Posición falsa

Para encontrar una solución para $f(x) = 0$ dada la función f continua en el intervalo $[p_0, p_1]$ donde $f(p_0)$ y $f(p_1)$ tienen signos opuestos:

ENTRADA aproximaciones iniciales p_0, p_1 tolerancia TOL; número máximo de iteraciones N_0.

SALIDA solución aproximada p o mensaje de falla.

Paso 1 Determine $i = 2$;
$q_0 = f(p_0)$;
$q_1 = f(p_1)$.

Paso 2 Mientras $i \leq N_0$ haga los pasos 3–7.

Paso 3 Determine $p = p_1 - q_1(p_1 - p_0)/(q_1 - q_0)$. (*Calcule* p_i.)

Paso 4 Si $|p - p_1| < TOL$ entonces
SALIDA (p); (*El procedimiento fue exitoso*.)
PARE.

Paso 5 Determine $i = i + 1$;
$q = f(p)$.

Paso 6 Si $q \cdot q_1 < 0$ entonces determine $p_0 = p_1$;
$q_0 = q_1$.

Paso 7 Determine $p_1 = p$;
$q_1 = q$.

Paso 8 SALIDA ('El método falló después de N_0 iteraciones, $N_0 =$', N_0);
(*El procedimiento no fue exitoso*.)
PARE. ■

Ejemplo 3 Use el método de posición falsa para encontrar una solución a $x = \cos x$ y compare las aproximaciones con aquellas determinadas en el ejemplo 1, que aplican la iteración del punto fijo y el método de Newton, y con aquellas encontradas en el ejemplo 2, que aplica el método de la secante.

Solución Para efectuar una comparación razonable usaremos las mismas aproximaciones iniciales que en el método de la secante; es decir $p_0 = 0.5$ y $p_1 = \pi/4$. La tabla 2.6 muestra los resultados del método de posición falsa aplicado a $f(x) = \cos x - x$ junto con los obtenidos mediante los métodos de la secante y de Newton. Observe que las aproximaciones de posición falsa y de la secante concuerdan a través de p_3 y que el método de posición falsa requiere una iteración adicional para obtener la misma precisión que el método de la secante. ■

Tabla 2.6

	Posición falsa	Secante	Newton
n	p_n	p_n	p_n
0	0.5	0.5	0.7853981635
1	0.7853981635	0.7853981635	0.7395361337
2	0.7363841388	0.7363841388	0.7390851781
3	0.7390581392	0.7390581392	0.7390851332
4	0.7390848638	0.7390851493	0.7390851332
5	0.7390851305	0.7390851332	
6	0.7390851332		

Comúnmente, la seguridad adicional del método de posición falsa requiere más cálculos que el método de la secante, de la misma forma en que la simplificación proporcionada por el método de la secante sobre el método de Newton se realiza a expensas de iteraciones adicionales. Más ejemplos de las características positivas y negativas de estos métodos se pueden observar al trabajar en los ejercicios 13 y 14.

La sección Conjunto de ejercicios 2.3 está disponible en línea. Encuentre la ruta de acceso en las páginas preliminares.

2.4 Análisis de error para métodos iterativos

En esta sección investigamos el orden de convergencia de esquemas de iteración funcional y, con el propósito de obtener convergencia rápida, redescubrimos el método de Newton. También consideramos formas para acelerar la convergencia del método de Newton en circunstancias especiales. Primero, sin embargo, necesitamos un nuevo procedimiento para medir qué tan rápido converge una sucesión.

Orden de convergencia

Definición 2.7 Suponga que $\{p_n\}_{n=0}^{\infty}$ es una sucesión que converge a p, con $p_n \neq p$ para todas las n. Si existen constantes positivas λ y α con

$$\lim_{n\to\infty} \frac{|p_{n+1} - p|}{|p_n - p|^{\alpha}} = \lambda,$$

Entonces $\{p_n\}_{n=0}^{\infty}$ **converge a p de orden α, con constante de error asintótica λ.** ■

Se dice que una técnica iterativa de la forma $p_n = g(p_{n-1})$ es de *orden* α si la sucesión $\{p_n\}_{n=0}^{\infty}$ converge a la solución $p = g(p)$ de orden α.

En general, una sucesión con un alto orden converge más rápidamente que una sucesión con un orden más bajo. La constante asintótica afecta la velocidad de convergencia pero no el grado del orden. Se presta atención especial a dos casos:

i) Si $\alpha = 1$ (y $\lambda < 1$), la sucesión es **linealmente convergente**.

ii) Si $\alpha = 2$, la sucesión es **cuadráticamente convergente**.

La siguiente ilustración compara una linealmente convergente con una que es cuadráticamente convergente. Y se muestra por qué tratamos de encontrar métodos que producen sucesiones convergentes de orden superior.

Ilustración Suponga que $\{p_n\}_{n=0}^{\infty}$ es linealmente convergente en 0 con

$$\lim_{n\to\infty} \frac{|p_{n+1}|}{|p_n|} = 0.5$$

Y que $\{\tilde{p}_n\}_{n=0}^{\infty}$ es cuadráticamente convergente a 0 con la misma constante de error asintótico,

$$\lim_{n\to\infty} \frac{|\tilde{p}_{n+1}|}{|\tilde{p}_n|^2} = 0.5.$$

Por simplicidad, asumimos que para cada n, tenemos

$$\frac{|p_{n+1}|}{|p_n|} \approx 0.5 \quad \text{y} \quad \frac{|\tilde{p}_{n+1}|}{|\tilde{p}_n|^2} \approx 0.5.$$

Para el esquema linealmente convergente, esto significa que

$$|p_n - 0| = |p_n| \approx 0.5|p_{n-1}| \approx (0.5)^2|p_{n-2}| \approx \cdots \approx (0.5)^n|p_0|,$$

mientras el procedimiento cuadráticamente convergente es

$$\begin{aligned}|\tilde{p}_n - 0| = |\tilde{p}_n| &\approx 0.5|\tilde{p}_{n-1}|^2 \approx (0.5)[0.5|\tilde{p}_{n-2}|^2]^2 = (0.5)^3|\tilde{p}_{n-2}|^4\\ &\approx (0.5)^3[(0.5)|\tilde{p}_{n-3}|^2]^4 = (0.5)^7|\tilde{p}_{n-3}|^8\\ &\approx \cdots \approx (0.5)^{2^n-1}|\tilde{p}_0|^{2^n}.\end{aligned}$$

La tabla 2.7 ilustra la velocidad relativa de convergencia de las sucesiones en 0 si $|p_0| = |\tilde{p}_0| = 1$.

Tabla 2.7

n	Sucesión de convergencia lineal $\{p_n\}_{n=0}^{\infty}$ $(0.5)^n$	Sucesión de convergencia cuadrática $\{\tilde{p}_n\}_{n=0}^{\infty}$ $(0.5)^{2^n-1}$
1	5.0000×10^{-1}	5.0000×10^{-1}
2	2.5000×10^{-1}	1.2500×10^{-1}
3	1.2500×10^{-1}	7.8125×10^{-3}
4	6.2500×10^{-2}	3.0518×10^{-5}
5	3.1250×10^{-2}	4.6566×10^{-10}
6	1.5625×10^{-2}	1.0842×10^{-19}
7	7.8125×10^{-3}	5.8775×10^{-39}

La sucesión cuadráticamente convergente se encuentra dentro de 10^{-38} de 0 mediante el séptimo término. Se necesitan por lo menos 126 términos para garantizar esta precisión para la sucesión linealmente convergente. ■

Se espera que las sucesiones cuadráticamente convergentes converjan mucho más rápido que las que sólo convergen linealmente, pero el siguiente resultado implica que una técnica de punto fijo arbitraria que genera secuencias convergentes sólo lo hace linealmente.

Teorema 2.8 Sea $g \in [a, b]$ tal que $g(x) \in [a, b]$ para todas las $x \in [a, b]$. Suponga además que g' es continua en (a, b) y que existe una constante positiva $k < 1$ con

$$|g'(x)| \leq k, \quad \text{para toda } x \in (a, b).$$

Si $g'(p) \neq 0$, entonces para cualquier número $p_0 \neq p$ en $[a, b]$, la sucesión

$$p_n = g(p_{n-1}), \quad \text{para } n \geq 1,$$

Converge sólo linealmente para el único punto fijo p en $[a, b]$.

Demostración Sabemos que, a partir del teorema de punto fijo en la sección 2.2, la sucesión converge a p. Puesto que existe g' en (a, b), podemos aplicar el teorema del valor medio para g para demostrar que para cualquier n,

$$p_{n+1} - p = g(p_n) - g(p) = g'(\xi_n)(p_n - p),$$

donde ξ_n está entre p_n y p. Ya que $\{p_n\}_{n=0}^{\infty}$ converge a p, también tenemos que $\{\xi_n\}_{n=0}^{\infty}$ converge a p. Puesto que g_0 es continua en (a, b), tenemos

$$\lim_{n\to\infty} g'(\xi_n) = g'(p).$$

Por lo tanto

$$\lim_{n\to\infty} \frac{p_{n+1}-p}{p_n-p} = \lim_{n\to\infty} g'(\xi_n) = g'(p) \quad \text{y} \quad \lim_{n\to\infty} \frac{|p_{n+1}-p|}{|p_n-p|} = |g'(p)|.$$

De este modo, si $g'(p) \neq 0$, la iteración de punto fijo muestra convergencia lineal con error asintótico constante $|g'(p)|$. ■

El teorema 2.8 implica que la convergencia de orden superior para los métodos de punto fijo de la forma $g(p) = p$ sólo se puede presentar cuando $g'(p) = 0$. El siguiente resultado describe condiciones adicionales que garantizan la convergencia cuadrática que buscamos.

Teorema 2.9 Sea p una solución de la ecuación $x = g(x)$. Suponga que $g'(p) = 0$ y que g'' es continua con $|g''(x)| < M$ en un intervalo abierto I que contiene a p. Entonces existe $\delta > 0$ tal que para $p_0 \in [p-\delta, p+\delta]$, la sucesión definida por $p_n = g(p_{n-1})$, cuando $n \geq 1$, converge, por lo menos cuadráticamente a p. Además, con valores suficientemente grandes de n,

$$|p_{n+1}-p| < \frac{M}{2}|p_n-p|^2.$$

Demostración Seleccione k en $(0, 1)$ y $\delta > 0$ tal que en el intervalo $[p-\delta, p+\delta]$, contenido en I, tenemos $|g'(x)| \leq k$ y g'' continua. Puesto que $|g'(x)| \leq k < 1$, el argumento utilizado en la prueba del teorema 2.6 en la sección 2.3 muestra que los términos de la sucesión $\{p_n\}_{n=0}^{\infty}$ están contenidos en $[p-\delta, p+\delta]$. Al expandir $g(x)$ en un polinomio lineal de Taylor, para $x \in [p-\delta, p+\delta]$ obtenemos

$$g(x) = g(p) + g'(p)(x-p) + \frac{g''(\xi)}{2}(x-p)^2,$$

donde ξ se encuentra entre x y p. Las hipótesis $g(p) = p$ y $g'(p) = 0$ implican que

$$g(x) = p + \frac{g''(\xi)}{2}(x-p)^2.$$

En especial, cuando $x = p_n$,

$$p_{n+1} = g(p_n) = p + \frac{g''(\xi_n)}{2}(p_n-p)^2,$$

con ξ_n entre p_n y p. Por lo tanto,

$$p_{n+1} - p = \frac{g''(\xi_n)}{2}(p_n-p)^2.$$

Puesto que $|g'(x)| \leq k < 1$ en $[p-\delta, p+\delta]$ y g mapea $[p-\delta, p+\delta]$ en sí mismo, por el teorema de punto fijo se sigue que $\{p_n\}_{n=0}^{\infty}$ converge a p. Pero como ξ_n se encuentra entre p y p_n para cada n, entonces $\{\xi_n\}_{n=0}^{\infty}$ también convergen a p y

$$\lim_{n\to\infty} \frac{|p_{n+1}-p|}{|p_n-p|^2} = \frac{|g''(p)|}{2}.$$

Este resultado implica que la sucesión $\{p_n\}_{n=0}^{\infty}$ es cuadráticamente convergente si $g''(p) \neq 0$ y de convergencia de orden superior si $g''(p) = 0$.

Puesto que g'' es continua y está estrictamente acotada por M en el intervalo $[p-\delta, p+\delta]$, esto también implica que, para los valores suficientemente grandes de n,

$$|p_{n+1}-p| < \frac{M}{2}|p_n-p|^2.$$

■

Los teoremas 2.8 y 2.9 nos indican que nuestra búsqueda de métodos de punto fijo que convergen cuadráticamente deberían señalar hacia funciones cuyas derivadas son cero en el punto fijo. Es decir,

- Para un método de punto fijo que converge cuadráticamente, necesitamos tener tanto $g(p) = p$ como $g'(p) = 0$.

La forma más fácil de construir un problema de punto fijo relacionado con el problema de encontrar la raíz $f(x) = 0$ es sumar o restar un múltiplo de $f(x)$ a partir de x. Considere la sucesión

$$p_n = g(p_{n-1}), \quad \text{para } n \geq 1,$$

para g en la forma de

$$g(x) = x - \phi(x) f(x),$$

donde ϕ es una función diferenciable que se seleccionará más adelante.

Para que el procedimiento iterativo derivado de g sea convergente cuadráticamente, necesitamos tener $g'(p) = 0$ cuando $f(p) = 0$. Ya que

$$g'(x) = 1 - \phi'(x) f(x) - f'(x)\phi(x)$$

y $f(p) = 0$, tenemos

$$g'(p) = 1 - \phi'(p) f(p) - f'(p)\phi(p) = 1 - \phi'(p) \cdot 0 - f'(p)\phi(p) = 1 - f'(p)\phi(p),$$

y $g'(p) = 0$ si y sólo si $\phi(p) = 1/f'(p)$.

Si hacemos $\phi(x) = 1/f'(x)$ entonces garantizamos que el procedimiento converge cuadráticamente

$$p_n = g(p_{n-1}) = p_{n-1} - \frac{f(p_{n-1})}{f'(p_{n-1})}.$$

Esto, por supuesto, es simplemente el método de Newton. Por lo tanto,

- Si $f(p) = 0$ y $f'(p) \neq 0$, entonces para los valores suficientemente cercanos a p, el método de Newton convergerá por lo menos cuadráticamente.

Raíces múltiples

En el análisis anterior se efectuó una restricción en la que $f'(p) \neq 0$, donde p es la solución de $f(x) = 0$. En especial, el método de Newton y el método de la secante, en general, dan problemas si $f'(p) = 0$ cuando $f(p) = 0$. Para examinar estas dificultades más detalladamente, damos la siguiente definición.

Definición 2.10 Una solución p de $f(x) = 0$ es un **cero de multiplicidad** m de f si para $x \neq p$, podemos escribir $f(x) = (x - p)^m q(x)$, donde $\lim_{x \to p} q(x) \neq 0$. ■

Para polinomios, p es un cero de multiplicidad m de f, si $f(x) = (x - p)^m q(x)$, donde $q(p) \neq 0$.

En esencia, $q(x)$ representa la porción de $f(x)$ que no contribuye con el cero de f. El siguiente resultado proporciona los medios para identificar fácilmente los ceros **simples** de una función, aquellos que tienen multiplicidad uno.

Teorema 2.11 La función $f \in C^1[a, b]$ tiene un cero simple en p en (a, b) si y sólo si $f(p) = 0$, pero $f'(p) \neq 0$.

Demostración Si f tiene un cero simple en p, entonces $f(p) = 0$ y $f(x) = (x - p)q(x)$, donde $\lim_{x\to p} q(x) \neq 0$. Ya que $f \in C^1[a, b]$,

$$f'(p) = \lim_{x\to p} f'(x) = \lim_{x\to p}[q(x) + (x - p)q'(x)] = \lim_{x\to p} q(x) \neq 0.$$

Por otra parte, si $f(p) = 0$ pero $f'(p) \neq 0$, represente a f como polinomio de Taylor de grado cero alrededor de p. Entonces

$$f(x) = f(p) + f'(\xi(x))(x - p) = (x - p)f'(\xi(x)),$$

donde $\xi(x)$ está entre x y p. Ya que $f \in C^1[a, b]$.

$$\lim_{x\to p} f'(\xi(x)) = f'\left(\lim_{x\to p} \xi(x)\right) = f'(p) \neq 0.$$

Haciendo $q = f' \circ \xi$ obtenemos $f(x) = (x - p)q(x)$, donde $\lim_{x\to p} q(x) \neq 0$. Por lo tanto, f tiene un cero simple en p. ■

La siguiente generalización del teorema 2.11 se considera en el ejercicio 12.

Teorema 2.12 La función $f \in C^m[a, b]$ tiene un cero de multiplicidad m en p en (a, b) si y sólo si

$$0 = f(p) = f'(p) = f''(p) = \cdots = f^{(m-1)}(p), \quad \text{pero} \quad f^{(m)}(p) \neq 0.$$ ■

El resultado en el teorema 2.12 implica que existe un intervalo alrededor de p donde el método de Newton converge cuadráticamente a p para cualquier aproximación inicial $p_0 = p$, siempre que p sea un cero simple. El siguiente ejemplo muestra que la convergencia cuadrática podría no presentarse si el cero no es simple.

Ejemplo 1 Sea $f(x) = e^x - x - 1$. **a)** Muestre que f tiene un cero de multiplicidad 2 en $x = 0$. **b)** Muestre que el método de Newton con $p_0 = 1$ converge para este cero, pero no de manera cuadrática.

Solución **a)** Tenemos

$$f(x) = e^x - x - 1, \qquad f'(x) = e^x - 1, \qquad \text{y} \qquad f''(x) = e^x,$$

por lo que

$$f(0) = e^0 - 0 - 1 = 0, \qquad f'(0) = e^0 - 1 = 0, \qquad \text{y} \qquad f''(0) = e^0 = 1.$$

El teorema 2.12 implica que f tiene un cero de multiplicidad 2 en $x = 0$.
b) Los primeros dos términos generados por el método de Newton aplicado a f con $p_0 = 1$ son

$$p_1 = p_0 - \frac{f(p_0)}{f'(p_0)} = 1 - \frac{e-2}{e-1} \approx 0.58198$$

y

$$p_2 = p_1 - \frac{f(p_1)}{f'(p_1)} \approx 0.58198 - \frac{0.20760}{0.78957} \approx 0.31906.$$

Los primeros ocho términos de la sucesión generada por el método de Newton se muestran en la tabla 2.8. La sucesión es claramente convergente a 0, pero no cuadráticamente. La gráfica de f se muestra en la figura 2.11. ■

Figura 2.11

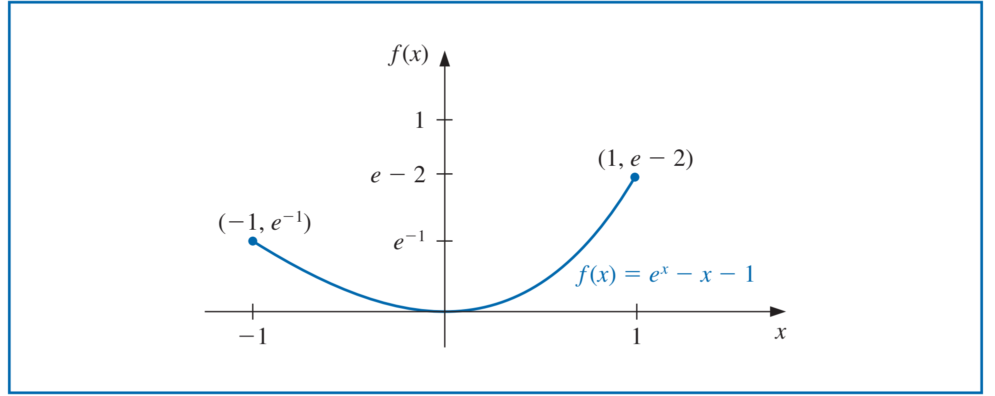


Tabla 2.8

n	p_n
0	1.0
1	0.58198
2	0.31906
3	0.16800
4	0.08635
5	0.04380
6	0.02206
7	0.01107
8	0.005545
9	2.7750×10^{-3}
10	1.3881×10^{-3}
11	6.9411×10^{-4}
12	3.4703×10^{-4}
13	1.7416×10^{-4}
14	8.8041×10^{-5}
15	4.2610×10^{-5}
16	1.9142×10^{-6}

Un método para manejar el problema de raíces múltiples de una función f es definir

$$\mu(x) = \frac{f(x)}{f'(x)}.$$

Si p es un cero de f de multiplicidad m con $f(x) = (x-p)^m q(x)$, entonces

$$\begin{aligned}\mu(x) &= \frac{(x-p)^m q(x)}{m(x-p)^{m-1}q(x) + (x-p)^m q'(x)} \\ &= (x-p)\frac{q(x)}{mq(x) + (x-p)q'(x)}\end{aligned}$$

también tiene un cero en p. Sin embargo, $q(p) \neq 0$, por lo que

$$\frac{q(p)}{mq(p) + (p-p)q'(p)} = \frac{1}{m} \neq 0,$$

y p es un cero simple de $\mu(x)$. Entonces, el método de Newton se puede aplicar a $\mu(x)$ para proporcionar

$$g(x) = x - \frac{\mu(x)}{\mu'(x)} = x - \frac{f(x)/f'(x)}{\{[f'(x)]^2 - [f(x)][f''(x)]\}/[f'(x)]^2},$$

Lo cual se simplifica en

$$g(x) = x - \frac{f(x)f'(x)}{[f'(x)]^2 - f(x)f''(x)}. \tag{2.13}$$

Si g tiene las condiciones de continuidad requeridas, la iteración funcional aplicada a g será convergente cuadráticamente, sin importar la multiplicidad del cero de f. En teoría el único inconveniente de este método es el cálculo adicional de $f''(x)$ y el procedimiento más laborioso que consiste en calcular las iteraciones. Sin embargo, en la práctica las raíces múltiples pueden causar graves problemas de redondeo debido a que el denominador de la ecuación (2.13) consiste en la diferencia de dos números, ambos cerca de 0.

Ejemplo 2 En el ejemplo 1 se mostró que $f(x) = e^x - x - 1$ tiene un cero de multiplicidad 2 en $x = 0$ y que el método de Newton con $p_0 = 1$ converge en este cero, pero no de manera cuadrática. Muestre que la modificación del método de Newton, como se indica en la ecuación (2.13) mejora la tasa de convergencia.

Tabla 2.9

n	p_n
1	$-2.3421061 \times 10^{-1}$
2	$-8.4582788 \times 10^{-3}$
3	$-1.1889524 \times 10^{-5}$
4	$-6.8638230 \times 10^{-6}$
5	$-2.8085217 \times 10^{-7}$

Solución El método modificado de Newton da

$$p_1 = p_0 - \frac{f(p_0)f'(p_0)}{f'(p_0)^2 - f(p_0)f''(p_0)} = 1 - \frac{(e-2)(e-1)}{(e-1)^2 - (e-2)e} \approx -2.3421061 \times 10^{-1}.$$

Esto está considerablemente más cerca de 0 que el primer término al usar el método de Newton, que era 0.58918. La tabla 2.9 enumera las primeras cinco aproximaciones para el doble cero en $x = 0$. Los resultados se obtuvieron a partir de un sistema con 10 dígitos de precisión. La falta relativa de mejora en las últimas dos entradas se debe al hecho de que al usar este sistema, tanto el numerador como el denominador se aproximan a 0. Por consiguiente, existe una pérdida de dígitos significativos de precisión, conforme las aproximaciones se acercan a 0.

Lo siguiente ilustra que el método modificado de Newton converge cuadráticamente incluso en el caso de un cero simple.

Ilustración En la sección 2.2, encontramos que un cero de $f(x) = x^3 + 4x^2 - 10 = 0$ es $p = 1.36523001$. Aquí compararemos la convergencia de un cero simple usando tanto el método de Newton como el método modificado de Newton, dado en la ecuación (2.13). Sea

i) $$p_n = p_{n-1} - \frac{p_{n-1}^3 + 4p_{n-1}^2 - 10}{3p_{n-1}^2 + 8p_{n-1}},$$ a partir del método de Newton,

y a partir del método modificado de Newton provisto por la ecuación (2.13),

ii) $$p_n = p_{n-1} - \frac{(p_{n-1}^3 + 4p_{n-1}^2 - 10)(3p_{n-1}^2 + 8p_{n-1})}{(3p_{n-1}^2 + 8p_{n-1})^2 - (p_{n-1}^3 + 4p_{n-1}^2 - 10)(6p_{n-1} + 8)}.$$

Con $p_0 = 1.5$, tenemos

Método de Newton

$$p_1 = 1.37333333, \quad p_2 = 1.36526201, \quad \text{y} \quad p_3 = 1.36523001.$$

Método modificado de Newton

$$p_1 = 1.35689898, \quad p_2 = 1.36519585, \quad \text{y} \quad p_3 = 1.36523001.$$

Ambos métodos son rápidamente convergentes al cero real, el cual es dado por ambos métodos como p_3. Sin embargo, observe que, en el caso de un cero simple, el método original de Newton requiere considerablemente menos cálculos. ■

La sección Conjunto de ejercicios 2.4 está disponible en línea. Encuentre la ruta de acceso en las páginas preliminares.

2.5 Convergencia acelerada

El teorema 2.8 indica que es raro tener el lujo de la convergencia cuadrática. Ahora consideramos una técnica llamada **método Δ^2 de Aitken**, que se puede utilizar para acelerar la convergencia de una sucesión que es linealmente convergente, independientemente de su origen o aplicación.

Método Δ^2 de Aitken

Suponga $\{p_n\}_{n=0}^{\infty}$ es una sucesión linealmente convergente con límite p. Para motivar la construcción de una sucesión $\{\hat{p}_n\}_{n=0}^{\infty}$ que converge más rápidamente a p que $\{p_n\}_{n=0}^{\infty}$

Alexander Aitken (1895–1967) usó esta técnica en 1926 para acelerar la tasa de convergencia de una serie en un artículo sobre ecuaciones algebraicas [Ai]. Este proceso es similar al que usó mucho antes el matemático japonés Takakazu Seki Kowa (1642–1708).

suponga que los signos de $p_n - p$, $p_{n+1} - p$, y $p_{n+2} - p$ concuerdan y que n es suficientemente grande para que

$$\frac{p_{n+1} - p}{p_n - p} \approx \frac{p_{n+2} - p}{p_{n+1} - p}.$$

Entonces

$$(p_{n+1} - p)^2 \approx (p_{n+2} - p)(p_n - p),$$

por lo que

$$p_{n+1}^2 - 2p_{n+1}p + p^2 \approx p_{n+2}p_n - (p_n + p_{n+2})p + p^2$$

y

$$(p_{n+2} + p_n - 2p_{n+1})p \approx p_{n+2}p_n - p_{n+1}^2.$$

Al resolver p obtenemos

$$p \approx \frac{p_{n+2}p_n - p_{n+1}^2}{p_{n+2} - 2p_{n+1} + p_n}.$$

Al sumar y restar los términos p_n^2 y $2p_np_{n+1}$ en el numerador y agrupar los términos adecuadamente obtenemos

$$\begin{aligned} p &\approx \frac{p_np_{n+2} - 2p_np_{n+1} + p_n^2 - p_{n+1}^2 + 2p_np_{n+1} - p_n^2}{p_{n+2} - 2p_{n+1} + p_n} \\ &= \frac{p_n(p_{n+2} - 2p_{n+1} + p_n) - (p_{n+1}^2 - 2p_np_{n+1} + p_n^2)}{p_{n+2} - 2p_{n+1} + p_n} \\ &= p_n - \frac{(p_{n+1} - p_n)^2}{p_{n+2} - 2p_{n+1} + p_n}. \end{aligned}$$

El **método Δ^2 de Aitken** se basa en la suposición de que la sucesión definida por $\{\hat{p}_n\}_{n=0}^{\infty}$.

$$\hat{p}_n = p_n - \frac{(p_{n+1} - p_n)^2}{p_{n+2} - 2p_{n+1} + p_n}, \quad (2.14)$$

converge más rápidamente a p que la sucesión original $\{p_n\}_{n=0}^{\infty}$.

Ejemplo 1 La sucesión $\{p_n\}_{n=1}^{\infty}$, donde $p_n = \cos(1/n)$, converge linealmente a $p = 1$. Determine los primeros cinco términos de la sucesión provista por el método Δ^2 de Aitken.

Tabla 2.10

n	p_n	$\hat{p}_n$
1	0.54030	0.96178
2	0.87758	0.98213
3	0.94496	0.98979
4	0.96891	0.99342
5	0.98007	0.99541
6	0.98614	
7	0.98981	

Solución Con el fin de determinar el término $\hat{p}_n$ de la sucesión con el método Δ^2 de Aitken, necesitamos tener los términos p_n, p_{n+1}, y p_{n+2} de la sucesión original. Por lo tanto, para determinar $\hat{p}_5$, necesitamos los primeros siete términos de $\{p_n\}$. Estos se muestran en la tabla 2.10. Ciertamente, parece que $\{\hat{p}_n\}_{n=1}^{\infty}$ converge más rápido a $p = 1$ que $\{p_n\}_{n=1}^{\infty}$. ■

La notación Δ relacionada con esta técnica tiene su origen en la siguiente definición.

Definición 2.13 Para una sucesión $\{p_n\}_{n=0}^{\infty}$ determinada, la **diferencia hacia adelante** Δp_n (que se lee "delta p_n") se define mediante

$$\Delta p_n = p_{n+1} - p_n, \quad \text{para } n \geq 0.$$

Las potencias superiores del operador Δ se definen de manera recursiva con

$$\Delta^k p_n = \Delta(\Delta^{k-1} p_n), \quad \text{para } k \geq 2.$$ ■

La definición implica que

$$\Delta^2 p_n = \Delta(p_{n+1} - p_n) = \Delta p_{n+1} - \Delta p_n = (p_{n+2} - p_{n+1}) - (p_{n+1} - p_n).$$

Por lo que $\Delta^2 p_n = p_{n+2} - 2p_{n+1} + p_n$, y la fórmula para $\hat{p}_n$ determinada en la ecuación (2.14) se puede escribir como

$$\hat{p}_n = p_n - \frac{(\Delta p_n)^2}{\Delta^2 p_n}, \quad \text{para } n \geq 0. \tag{2.15}$$

En este punto de nuestro análisis del método Δ^2 de Aitken, hemos establecido que la sucesión $\{\hat{p}_n\}_{n=0}^{\infty}$ converge a p más rápidamente que la sucesión original $\{p_n\}_{n=0}^{\infty}$, pero no hemos dicho lo que significa el término convergencia "más rápida". El teorema 2.14 explica y justifica esta terminología. La prueba de este teorema se considera en el ejercicio 16.

Teorema 2.14 Suponga que $\{p_n\}_{n=0}^{\infty}$, es una sucesión que converge linealmente en el límite p y que

$$\lim_{n\to\infty} \frac{p_{n+1} - p}{p_n - p} < 1.$$

Entonces la secuencia Δ^2 de Aitken $\{\hat{p}_n\}_{n=0}^{\infty}$ converge a p más rápido que $\{p_n\}_{n=0}^{\infty}$ en el sentido en que

$$\lim_{n\to\infty} \frac{\hat{p}_n - p}{p_n - p} = 0.$$ ■

Método de Steffensen

Johan Frederik Steffensen (1873–1961) escribió un prestigioso libro titulado *Interpolation* en 1927.

Al aplicar una modificación del método Δ^2 de Aitken a una sucesión linealmente convergente obtenida a partir de la iteración de punto fijo, podemos acelerar la convergencia a cuadrática. Este procedimiento recibe el nombre de método de Steffensen y difiere ligeramente de la aplicación del método Δ^2 de Aitken directamente para la sucesión de iteración de punto fijo linealmente convergente. El método Δ^2 de Aitken construye los términos en el orden:

$$p_0, \quad p_1 = g(p_0), \quad p_2 = g(p_1), \quad \hat{p}_0 = \{\Delta^2\}(p_0),$$

$$p_3 = g(p_2), \quad \hat{p}_1 = \{\Delta^2\}(p_1), \ldots,$$

donde $\{\Delta^2\}$ indica que se usa la ecuación (2.15). El método de Steffensen construye los primeros cuatro términos p_0, p_1, p_2 y $\hat{p}_0$. Sin embargo, en este paso suponemos que $\hat{p}_0$ es una mejor aproximación a p que es p_2 y aplicamos la iteración de punto fijo a $\hat{p}_0$ en lugar de a p_2. Con esta notación, la sucesión generada es

$$p_0^{(0)}, \quad p_1^{(0)} = g(p_0^{(0)}), \quad p_2^{(0)} = g(p_1^{(0)}), \quad p_0^{(1)} = \{\Delta^2\}(p_0^{(0)}), \quad p_1^{(1)} = g(p_0^{(1)}), \ldots.$$

Cada tercer término de la sucesión de Steffensen se genera con la ecuación (2.15); los otros usan la iteración de punto fijo en el término previo. El proceso se describe en el algoritmo 2.6.

ALGORITMO 2.6

Método de Steffensen

Para encontrar una solución para $p = g(p)$ dada una aproximación p_0:

ENTRADA aproximación inicial p_0 tolerancia TOL; número máximo de iteraciones N_0.

SALIDA aproxime la solución p o mensaje de falla.

Paso 1 Determine $i = 1$.

Paso 2 Mientras $i \leq N_0$ haga los pasos 3–6.

Paso 3 Determine $p_1 = g(p_0)$; (*Calcule* $p_1^{(i-1)}$.)
$p_2 = g(p_1)$; (*Calcule* $p_2^{(i-1)}$.)
$p = p_0 - (p_1 - p_0)^2/(p_2 - 2p_1 + p_0)$. (*Calcule* $p_0^{(i)}$.)

Paso 4 Si $|p - p_0| < TOL$ entonces
SALIDA (p); (*Procedimiento completado exitosamente*.)
PARE.

Paso 5 Determine $i = i + 1$.

Paso 6 Determine $p_0 = p$. (*Actualice* p_0.)

Paso 7 SALIDA (‘El método falló después de N_0 iteraciones, $N_0 =$’, N_0);
(*Procedimiento no completado exitosamente*.)
PARE. ■

Observe que $\Delta^2 p_n$ podría ser 0, que introducirá un 0 en el denominador de la siguiente iteración. Si esto pasa, terminamos la sucesión y seleccionamos $p_2^{(n-1)}$ como la mejor aproximación.

Ilustración Para resolver $x^3 + 4x^2 - 10 = 0$ con el método de Steffensen, si $x^3 + 4x^2 = 10$, divida entre $x + 4$ y resuelva para x. Este procedimiento crea el método de punto fijo

$$g(x) = \left(\frac{10}{x+4}\right)^{1/2}.$$

Consideramos el método de punto fijo en la tabla 2.2, columna **d)** de la sección 2.2.

Al aplicar el procedimiento de Steffensen con $p_0 = 1.5$ obtenemos los valores en la tabla 2.11. La iteración $p_0^{(2)} = 1.365230013$ es exacta hasta el noveno lugar decimal. En este ejemplo, el método de Steffensen ofrece casi la misma precisión que el método de Newton aplicado a este polinomio. Estos resultados se pueden observar en la ilustración al final de la sección 2.4. ■

Tabla 2.11

k	$p_0^{(k)}$	$p_1^{(k)}$	$p_2^{(k)}$
0	$p_0^{(0)}$	$p_1^{(0)} = g(p_0^{(0)})$	$p_2^{(0)} = g(p_1^{(0)})$
1	$p_0^{(1)} = p_0^{(0)} - \dfrac{(p_1^{(0)} - p_0^{(0)})^2}{p_2^{(0)} - 2p_1^{(0)} + p_0^{(0)}}$	$p_1^{(1)} = g(p_0^{(1)})$	$p_2^{(1)} = g(p_1^{(1)})$
2	$p_0^{(2)} = p_0^{(1)} - \dfrac{(p_1^{(1)} - p_0^{(1)})^2}{p_2^{(1)} - 2p_2^{(1)} + p_0^{(1)}}$		
Lo cual produce la siguiente tabla			
0	1.5	1.348399725	1.367376372
1	1.365265224	1.365225534	1.365230583
2	1.365230013		

A partir de la ilustración, parece que el método de Steffensen provee convergencia cuadrática sin evaluar una derivada y el teorema 2.14 establece que éste es el caso. La demostración de este teorema se puede encontrar en [He2], pp. 90–92 o [IK], pp. 103–107.

Teorema 2.15 Suponga que $x = g(x)$ tiene la solución p con $g'(p) \neq 1$. Si existe una $\delta > 0$ tal que $g \in C^3$ $[p - \delta, p + \delta]$, entonces el método de Steffensen proporciona convergencia cuadrática para cualquier $p_0 \in [p - \delta, p + \delta]$. ■

La sección Conjunto de ejercicios 2.5 está disponible en línea. Encuentre la ruta de acceso en las páginas preliminares.

2.6 Ceros de polinomios y método de Müller

Un *polinomio de grado n* tiene la forma

$$P(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0,$$

donde las a_i, llamadas *coeficientes* de P, son constantes y $a_n \neq 0$. La función cero $P(x) = 0$ para todos los valores de x se considera un polinomio, pero no tiene un grado asignado.

Polinomios algebraicos

Teorema 2.16 **(Teorema fundamental de álgebra)**

Si $P(x)$ es un polinomio de grado $n \geq 1$ con coeficientes reales o complejos, entonces $P(x) = 0$ tiene por lo menos una raíz (posiblemente compleja). ■

A pesar de que el teorema fundamental de álgebra es básico para cualquier estudio de las funciones elementales, la prueba usual requiere técnicas a partir del estudio de la teoría de función compleja. El lector puede consultar [SaS], p. 155, para la culminación de un desarrollo sistemático de los temas necesarios para mejorar este teorema.

Ejemplo 1 Determine todos los ceros del polinomio $P(x) = x^3 - 5x^2 + 17x - 13$.

Solución Es muy fácil verificar que $P(1) = 1 - 5 + 17 - 13 = 0$, por lo que $x = 1$ es un cero de P y $(x - 1)$ es un factor del polinomio. Al dividir $P(x)$ entre $x - 1$ obtenemos

$$P(x) = (x - 1)(x^2 - 4x + 13).$$

Para determinar los ceros de $x^2 - 4x + 13$, usamos la fórmula cuadrática en su forma estándar, la cual da los ceros complejos

$$\frac{-(-4) \pm \sqrt{(-4)^2 - 4(1)(13)}}{2(1)} = \frac{4 \pm \sqrt{-36}}{2} = 2 \pm 3i.$$

Por lo tanto, el polinomio de tercer grado $P(x)$ tiene tres ceros, $x_1 = 1$, $x_2 = 2 - 3i$ y $x_2 = 2 + 3i$. ■

Carl Friedrich Gauss (1777–1855), uno de los más grandes matemáticos de todos los tiempos, demostró el teorema fundamental de álgebra en su tesis doctoral y lo publicó en 1799. Publicó diferentes demostraciones de este resultado a lo largo de su vida, en 1815, 1816 y hasta en 1848. El resultado fue establecido, sin demostración, por Albert Girard (1595–1632), y Jean d'Alembert (1717–1783), Euler y Lagrange aportaron pruebas parciales.

En el ejemplo anterior, encontramos que el polinomio de tercer grado tiene tres ceros diferentes. Una consecuencia importante del teorema fundamental de álgebra es el siguiente corolario. Establece que éste siempre es el caso con la condición de que cuando los ceros no son distintos, contamos el número de ceros de acuerdo con sus multiplicidades.

Corolario 2.17 Si $P(x)$ es un polinomio de grado $n \geq 1$ con coeficientes reales o complejos, entonces existen constantes únicas $x_1, x_2, \ldots, x_k$, posiblemente complejas y enteros positivos únicos $m_1, m_2, \ldots, m_k$, tal que $\sum_{i=1}^{k} m_i = n$ y

$$P(x) = a_n(x - x_1)^{m_1}(x - x_2)^{m_2} \cdots (x - x_k)^{m_k}.$$

■

Mediante el corolario 2.17, la colección de ceros de un polinomio es única y, si cada cero x_i se cuenta tantas veces como su multiplicidad m_i, un polinomio de grado n tiene exactamente n ceros.

El siguiente corolario del teorema fundamental de álgebra se usa con frecuencia en esta sección y en capítulos posteriores.

Corolario 2.18 Sean $P(x)$ y $Q(x)$ polinomios de grado a lo más n. Si $x_1, x_2, \ldots, x_k$, con $k > n$, son números distintos con $P(x_i) = Q(x_i)$ para $i = 1, 2, \ldots, k$, entonces $P(x) = Q(x)$ para todos los valores de x. ■

Este resultado implica que para mostrar que dos polinomios de grado menor o igual que n son iguales, sólo necesitamos mostrar que concuerdan en $n + 1$ valores. Esto se usará con mucha frecuencia, especialmente en los capítulos 3 y 8.

William Horner (1786–1837) era un niño prodigio que se convirtió en maestro de una escuela en Bristol a los 18 años. El método de Horner para resolver ecuaciones algebraicas se publicó en 1819 en las *Philosophical Transactions of the Royal Society (Transacciones filosóficas de la Real Sociedad).*

Método de Horner

Al usar el método de Newton para localizar los ceros aproximados de un polinomio $P(x)$, necesitamos evaluar $P(x)$ y $P'(x)$ en valores específicos. Puesto que tanto $P(x)$ como $P'(x)$ son polinomios, la eficiencia computacional requiere que la evaluación de estas funciones se realice de la manera anidada que se analiza en la sección 1.2. El método de Horner incorpora esta técnica anidada y como consecuencia, sólo requiere n multiplicaciones y n sumas para evaluar un polinomio de enésimo grado.

Teorema 2.19 **(Método de Horner)**

Sea

$$P(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0.$$

Defina $b_n = a_n$ y

$$b_k = a_k + b_{k+1}x_0, \quad \text{para } k = n-1, n-2, \ldots, 1, 0.$$

Entonces $b_0 = P(x_0)$. Además, si

$$Q(x) = b_n x^{n-1} + b_{n-1}x^{n-2} + \cdots + b_2 x + b_1,$$

Entonces

$$P(x) = (x - x_0)Q(x) + b_0.$$

Paolo Ruffini (1765–1822) describió un método similar que lo hizo merecedor de la medalla de oro de la Italian Mathematical Society for Science (Sociedad Matemática Italiana para la Ciencia). Ni Ruffini ni Horner fueron los primeros en descubrir este método; ya se conocía en China 500 años antes.

Demostración Por la definición de $Q(x)$,

$$\begin{aligned}(x - x_0)Q(x) + b_0 &= (x - x_0)(b_n x^{n-1} + \cdots + b_2 x + b_1) + b_0\\ &= (b_n x^n + b_{n-1}x^{n-1} + \cdots + b_2 x^2 + b_1 x)\\ &\quad - (b_n x_0 x^{n-1} + \cdots + b_2 x_0 x + b_1 x_0) + b_0\\ &= b_n x^n + (b_{n-1} - b_n x_0)x^{n-1} + \cdots + (b_1 - b_2 x_0)x + (b_0 - b_1 x_0).\end{aligned}$$

Por la hipótesis, $b_n = a_n$ y $b_k - b_{k+1}x_0 = a_k$, por lo que

$$(x - x_0)Q(x) + b_0 = P(x) \quad \text{y} \quad b_0 = P(x_0).$$

■

Ejemplo 2 Use el método de Horner para evaluar $P(x) = 2x^4 - 3x^2 + 3x - 4$ en $x_0 = -2$.

Solución Cuando usamos el cálculo manual en el método de Horner, primero construimos una tabla que sugiere el nombre de la *división sintética*, que a menudo se aplica a esta técnica. Para este problema, la tabla aparece a continuación:

	Coeficiente de x^4	Coeficiente de x^3	Coeficiente de x^2	Coeficiente de x	Término constante
$x_0 = -2$	$a_4 = 2$	$a_3 = 0$	$a_2 = -3$	$a_1 = 3$	$a_0 = -4$
		$b_4x_0 = -4$	$b_3x_0 = 8$	$b_2x_0 = -10$	$b_1x_0 = 14$
	$b_4 = 2$	$b_3 = -4$	$b_2 = 5$	$b_1 = -7$	$b_0 = 10$

La palabra "sintético" tiene sus raíces en diferentes lenguajes. En inglés estándar, por lo general provee un sentido de algo que es "falso" o "sustituido". Sin embargo, en matemáticas toma la forma de algo que está "agrupado". La geometría sintética trata las formas como un todo en lugar de cómo objetos individuales, que es como se usa en geometría analítica. En la división sintética de polinomios, las diferentes potencias de las variables no se proporcionan de modo explícito, pero se mantienen juntas.

Por lo que,

$$P(x) = (x + 2)(2x^3 - 4x^2 + 5x - 7) + 10.$$ ■

Una ventaja adicional del uso del procedimiento de Horner (o de división sintética) es que, ya que

$$P(x) = (x - x_0)Q(x) + b_0,$$

donde

$$Q(x) = b_n x^{n-1} + b_{n-1}x^{n-2} + \cdots + b_2 x + b_1,$$

al diferenciar respecto a x obtenemos

$$P'(x) = Q(x) + (x - x_0)Q'(x) \quad \text{y} \quad P'(x_0) = Q(x_0). \tag{2.16}$$

Cuando el método de Newton-Raphson se usa para encontrar un cero aproximado de un polinomio, $P(x)$ y $P'(x)$ se puede evaluar de la misma forma.

Ejemplo 3 Encuentre una aproximación al cero de

$$P(x) = 2x^4 - 3x^2 + 3x - 4,$$

usando el método de Newton con $x_0 = -2$ y la división sintética para evaluar $P(x_n)$ y $P'(x_n)$ para cada iteración x_n.

Solución Con $x_0 = -2$ como aproximación inicial, obtuvimos $P(-2)$ en el ejemplo 1 mediante

$x_0 = -2$	2	0	−3	3	−4	
		−4	8	−10	14	
	2	−4	5	−7	10	$= P(-2)$.

Usando el teorema 2.19 y la ecuación (2.16),

$$Q(x) = 2x^3 - 4x^2 + 5x - 7 \quad \text{y} \quad P'(-2) = Q(-2),$$

por lo que $P'(-2)$ se puede encontrar al evaluar $Q(-2)$ en forma similar:

$x_0 = -2$	2	−4	5	−7	
		−4	16	−42	
	2	−8	21	−49	$= Q(-2) = P'(-2)$

y

$$x_1 = x_0 - \frac{P(x_0)}{P'(x_0)} = x_0 - \frac{P(x_0)}{Q(x_0)} = -2 - \frac{10}{-49} \approx -1.796.$$

Al repetir el procedimiento para encontrar x_2 obtenemos

-1.796	2	0	-3	3	-4	
		-3.592	6.451	-6.197	5.742	
	2	-3.592	3.451	-3.197	1.742	$= P(x_1)$
		-3.592	12.902	-29.368		
	2	-7.184	16.353	-32.565	$= Q(x_1)$	$= P'(x_1)$.

Por lo que, $P(-1.796) = 1.742$, $P'(-1.796) = Q(-1.796) = -32.565$, y

$$x_2 = -1.796 - \frac{1.742}{-32.565} \approx -1.7425.$$

De manera similar, $x_3 = -1.73897$, y un cero real con cinco cifras decimales es -1.73896.

Observe que el polinomio $Q(x)$ depende de la aproximación que se usa y cambia de una iteración a otra. ■

El algoritmo 2.7 calcula $P(x_0)$ y $P'(x_0)$ usando el método de Horner.

ALGORITMO 2.7

Método de Horner

Para evaluar el polinomio

$$P(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0 = (x - x_0)Q(x) + b_0$$

Y su derivada en x_0:

ENTRADA grado n; coeficientes $a_0, a_1, \ldots, a_n$; x_0.

SALIDA $y = P(x_0)$; $z = P'(x_0)$.

Paso 1 Determine $y = a_n$; (*Calcule b_n para P.*)
$z = a_n$. (*Calcule b_{n-1} para Q.*)

Paso 2 Para $j = n-1, n-2, \ldots, 1$
determine $y = x_0 y + a_j$; (*Calcule b_j para P.*)
$z = x_0 z + y$. (*Calcule b_{j-1} para Q.*)

Paso 3 Determine $y = x_0 y + a_0$. (*Calcule b_0 para P.*)

Paso 4 SALIDA (y, z);
PARE. ■

Si la enésima iteración, x_N, en el método de Newton es un cero aproximado para P, entonces

$$P(x) = (x - x_N)Q(x) + b_0 = (x - x_N)Q(x) + P(x_N) \approx (x - x_N)Q(x).$$

Por lo que, $x - x_N$ es un factor aproximado de $P(x)$. Suponiendo que $\hat{x}_1 = x_N$ sea el cero aproximado de P y $Q_1(x) \equiv Q(x)$ sea el factor aproximado obtenemos

$$P(x) \approx (x - \hat{x}_1)Q_1(x).$$

Podemos encontrar un segundo cero aproximado de P al aplicar el método de Newton para $Q_1(x)$.

Si $P(x)$ es un polinomio de enésimo grado con n ceros reales, este procedimiento aplicado repetidamente al final resultará en $(n-2)$ ceros aproximados de P y un factor cuadrático aproximado $Q_{n-2}(x)$. En esta etapa, $Q_{n-2}(x) = 0$ puede resolverse con la fórmula cuadrática para encontrar por lo menos dos ceros aproximados de P. A pesar de que este método se puede usar para encontrar todos los ceros aproximados, depende del uso repetido de aproximaciones y puede conducir a resultados imprecisos.

El procedimiento que se ha descrito recientemente recibe el nombre de **deflación**. La dificultad de la precisión con deflación se debe al hecho de que, cuando obtenemos los ceros aproximados de $P(x)$, el método de Newton se usa en el polinomio reducido $Q_k(x)$, es decir, el polinomio que tiene la propiedad de que

$$P(x) \approx (x - \hat{x}_1)(x - \hat{x}_2) \cdots (x - \hat{x}_k) Q_k(x).$$

Un cero aproximado $\hat{x}_{k+1}$ de Q_k por lo general no se aproximará a la raíz de $P(x) = 0$, como lo hace una raíz de la ecuación reducida $Q_k(x) = 0$ y la imprecisión aumenta conforme lo hace k. Una forma de eliminar esta dificultad es usar las ecuaciones reducidas para encontrar las aproximaciones $\hat{x}_2, \hat{x}_3, \ldots, \hat{x}_k$ para los ceros de P y, a continuación, mejorar estas aproximaciones al aplicar el método de Newton al polinomio original $P(x)$.

Ceros complejos: método de Müller

Un problema con la aplicación de los métodos de la secante, de posición falsa o de Newton a los polinomios, es la posibilidad de que el polinomio tenga raíces complejas incluso cuando todos los coeficientes son números reales. Si la aproximación inicial es un número real, todas las aproximaciones subsiguientes también serán números reales. Una forma de superar esta dificultad es comenzar con una aproximación inicial compleja y realizar todos los cálculos con aritmética compleja. Un enfoque alterno tiene sus bases en el siguiente teorema.

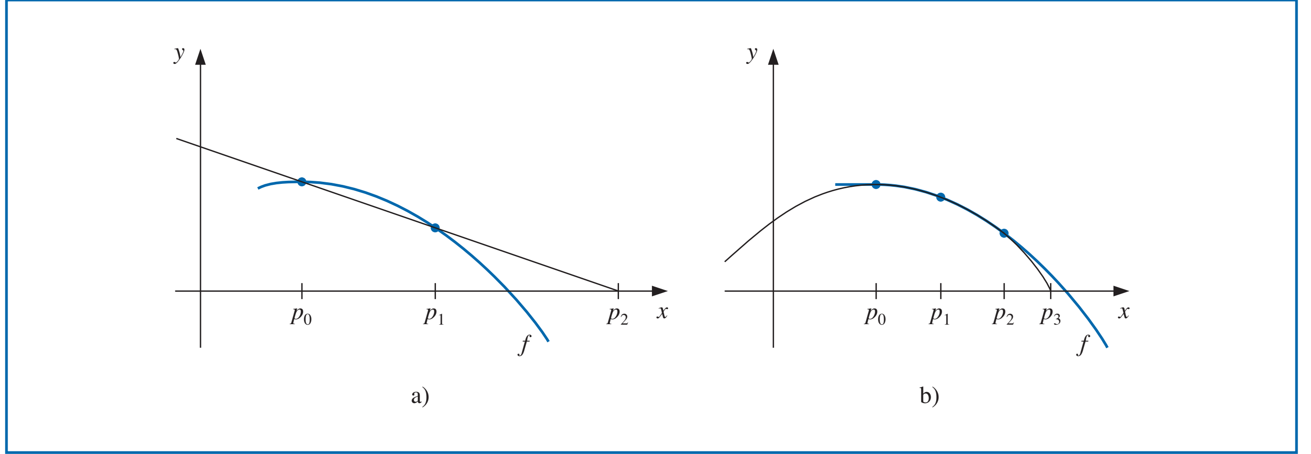
Teorema 2.20 Si $z = a + bi$ es un cero complejo de multiplicidad m del polinomio $P(x)$ con coeficientes reales, entonces $\bar{z} = a - bi$ también es un cero de multiplicidad m del polinomio $P(x)$ y $(x^2 - 2ax + a^2 + b^2)^m$ es un factor de $P(x)$. ■

Una división sintética mediante polinomios cuadráticos se puede concebir para factorizar aproximadamente el polinomio, de modo que un término será un polinomio cuadrático cuyas raíces complejas son aproximaciones a las raíces del polinomio original. Esta técnica se describió con cierto detalle en nuestra segunda edición [BFR]. En lugar de proceder con estas líneas, ahora consideraremos un método, que D. E. Müller [Mu] presentó primero. Esta técnica se puede usar para cualquier problema de encontrar la raíz, pero es especialmente útil para aproximar las raíces de los polinomios.

El método de Müller es similar al método de la secante. Sin embargo, mientras en el método de la secante se usa una recta que pasa por dos puntos en la curva para aproximar la raíz, en el método de Müller se utiliza una parábola a lo largo de tres puntos en la curva para la aproximación.

El método de la secante comienza con dos aproximaciones iniciales p_0 *y* p_1 y determina la siguiente aproximación p_2, como la intersección del eje x con la recta que pasa por $(p_0, f(p_0))$ y $(p_1, f(p_1))$. (Consulte la figura 2.12a).) En el método de Müller se usan tres aproximaciones iniciales, p_0, p_1 *y* p_2, y determina la siguiente aproximación p_3 al considerar la intersección del eje x con la parábola a través de $(p_0, f(p_0))$, $(p_1, f(p_1))$ y $(p_2, f(p_2))$. (Consulte la figura 2.12b).)

Figura 2.12



La derivación del método de Müller comienza al considerar el polinomio cuadrático

$$P(x) = a(x - p_2)^2 + b(x - p_2) + c$$

que pasa a través de $(p_0, f(p_0))$, $(p_1, f(p_1))$, y $(p_2, f(p_2))$. Las constantes a, b y c se pueden determinar a partir de las condiciones

$$f(p_0) = a(p_0 - p_2)^2 + b(p_0 - p_2) + c, \tag{2.17}$$

$$f(p_1) = a(p_1 - p_2)^2 + b(p_1 - p_2) + c, \tag{2.18}$$

y

$$f(p_2) = a \cdot 0^2 + b \cdot 0 + c = c \tag{2.19}$$

para ser

$$c = f(p_2), \tag{2.20}$$

$$b = \frac{(p_0 - p_2)^2[f(p_1) - f(p_2)] - (p_1 - p_2)^2[f(p_0) - f(p_2)]}{(p_0 - p_2)(p_1 - p_2)(p_0 - p_1)}, \tag{2.21}$$

y

$$a = \frac{(p_1 - p_2)[f(p_0) - f(p_2)] - (p_0 - p_2)[f(p_1) - f(p_2)]}{(p_0 - p_2)(p_1 - p_2)(p_0 - p_1)}. \tag{2.22}$$

Para determinar p_3, un cero de P, aplicamos la fórmula cuadrática para $P(x) = 0$. Sin embargo, debido a los problemas de error de redondeo causados por la resta de los números casi iguales, aplicamos la fórmula en la manera prescrita en las ecuaciones (1.2) y (1.3) de la sección 1.2:

$$p_3 - p_2 = \frac{-2c}{b \pm \sqrt{b^2 - 4ac}}.$$

Esta fórmula proporciona dos posibilidades para p_3, dependiendo del signo que precede al término radical. En el método de Müller, el signo se selecciona de acuerdo con el signo de b.

Elegido de esta forma, el denominador será el más grande en magnitud y resultará en p_3 siendo seleccionado como el cero más cercano de P para p_2. Por lo tanto,

$$p_3 = p_2 - \frac{2c}{b + \operatorname{sgn}(b)\sqrt{b^2 - 4ac}},$$

donde a, b y c se obtienen de las ecuaciones (2.20) a (2.22).

Una vez que se determina p_3, el procedimiento se reinicializa usando p_1, p_2 y p_3 en lugar de p_0, p_1 y p_2 para determinar la siguiente aproximación, p_4. El método continúa hasta obtener una conclusión satisfactoria. En cada paso, el método implica el radical $\sqrt{b^2 - 4ac}$, por lo que el método proporciona raíces complejas aproximadas cuando $b^2 - 4ac < 0$. El algoritmo 2.8 implementa este procedimiento.

ALGORITMO **2.8**

Método de Müller

Para encontrar una solución para $f(x) = 0$ dadas las tres aproximaciones p_0, p_1 y p_2:

ENTRADA p_0, p_1, p_2 tolerancia TOL; número máximo de iteraciones N_0.

SALIDA solución aproximada p o mensaje de falla.

Paso 1 Determine $h_1 = p_1 - p_0$;
$h_2 = p_2 - p_1$;
$\delta_1 = (f(p_1) - f(p_0))/h_1$;
$\delta_2 = (f(p_2) - f(p_1))/h_2$;
$d = (\delta_2 - \delta_1)/(h_2 + h_1)$;
$i = 3$.

Paso 2 Mientras $i \leq N_0$ haga los pasos 3–7.

Paso 3 $b = \delta_2 + h_2 d$;
$D = (b^2 - 4f(p_2)d)^{1/2}$. (*Nota: puede requerir aritmética compleja.*)

Paso 4 Si $|b - D| < |b + D|$ entonces determine $E = b + D$
también determine $E = b - D$.

Paso 5 Determine $h = -2f(p_2)/E$;
$p = p_2 + h$.

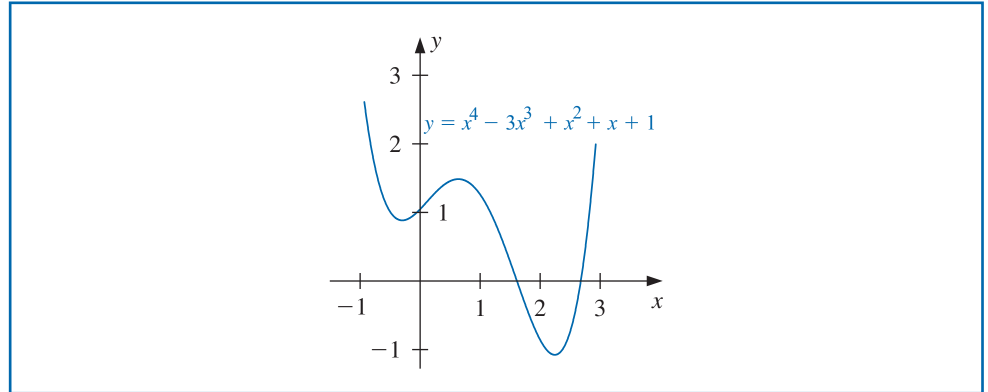
Paso 6 Si $|h| < TOL$ entonces
SALIDA (p); (*El procedimiento fue exitoso.*)
PARE.

Paso 7 Determine $p_0 = p_1$; (*Prepare la siguiente iteración.*)
$p_1 = p_2$;
$p_2 = p$;
$h_1 = p_1 - p_0$;
$h_2 = p_2 - p_1$;
$\delta_1 = (f(p_1) - f(p_0))/h_1$;
$\delta_2 = (f(p_2) - f(p_1))/h_2$;
$d = (\delta_2 - \delta_1)/(h_2 + h_1)$;
$i = i + 1$.

Paso 8 SALIDA ('El método falló después de N_0 iteraciones, N_0 =', N_0);
(*El procedimiento no fue exitoso.*)
PARE. ■

Ilustración Considere el polinomio $f(x) = x^4 - 3x^3 + x^2 + x + 1$, cuya gráfica se muestra en la figura 2.13.

Figura 2.13



Tres conjuntos de tres puntos iniciales se usarán con el algoritmo 2.8 y $TOL = 10^{-5}$ para aproximar los ceros de f. El primer conjunto usará $p_0 = 0.5$, $p_1 = -0.5$ y $p_2 = 0$. La parábola que pasa a través de estos puntos tiene raíces complejas porque no interseca el eje x. La tabla 2.12 provee aproximaciones para los ceros complejos correspondientes de f.

Tabla 2.12

	$p_0 = 0.5, \quad p_1 = -0.5, \quad p_2 = 0$	
i	p_i	$f(p_i)$
3	$-0.100000 + 0.888819i$	$-0.01120000 + 3.014875548i$
4	$-0.492146 + 0.447031i$	$-0.1691201 - 0.7367331502i$
5	$-0.352226 + 0.484132i$	$-0.1786004 + 0.0181872213i$
6	$-0.340229 + 0.443036i$	$0.01197670 - 0.0105562185i$
7	$-0.339095 + 0.446656i$	$-0.0010550 + 0.000387261i$
8	$-0.339093 + 0.446630i$	$0.000000 + 0.000000i$
9	$-0.339093 + 0.446630i$	$0.000000 + 0.000000i$

La tabla 2.13 nos da aproximaciones para los dos ceros reales de f. El más pequeño de éstos usa $p_0 = 0.5$, $p_1 = 1.0$ y $p_2 = 1.5$, y la raíz más grande se aproxima cuando $p_0 = 1.5$, $p_1 = 2.0$ y $p_2 = 2.5$.

Tabla 2.13

$p_0 = 0.5,\quad p_1 = 1.0,\quad p_2 = 1.5$			$p_0 = 1.5,\quad p_1 = 2.0,\quad p_2 = 2.5$		
i	p_i	$f(p_i)$	i	p_i	$f(p_i)$
3	1.40637	−0.04851	3	2.24733	−0.24507
4	1.38878	0.00174	4	2.28652	−0.01446
5	1.38939	0.00000	5	2.28878	−0.00012
6	1.38939	0.00000	6	2.28880	0.00000
			7	2.28879	0.00000

Los valores en las tablas son aproximaciones precisas para los lugares enumerados. ■

La ilustración muestra que el método de Müller puede aproximar las raíces de los polinomios con una variedad de valores iniciales. De hecho, el método de Müller por lo general converge con la raíz de un polinomio para cualquier selección de aproximación inicial, a pesar de que se pueden construir problemas para los cuales la convergencia no se presentará. Por ejemplo, suponga que para algunas i tenemos $f(p_i) = f(p_{i+1}) = f(p_{i+2}) \neq 0$. Entonces, la ecuación cuadrática se reduce a una función constante diferente de cero y nunca interseca el eje x. Sin embargo, esto no es normalmente el caso, y los paquetes de software de propósito general que usan el método de Müller sólo requieren una aproximación inicial por raíz e incluso proporcionarán esta aproximación como una opción.

La sección Conjunto de ejercicios 2.6 está disponible en línea. Encuentre la ruta de acceso en las páginas preliminares.



2.7 Software numérico y revisión del capítulo

Dada una función específica f y una tolerancia, un programa eficiente debería producir una aproximación para una o más soluciones de $f(x) = 0$, cada una tiene un error absoluto o relativo dentro de la tolerancia y los resultados deberían generarse en una cantidad razonable de tiempo. Si el programa no puede realizar esta tarea, por lo menos debería proporcionar explicaciones lógicas de porqué no se consiguió el éxito y una indicación sobre cómo remediar la causa de la falla.

IMSL tiene subrutinas que implementan el método de Müller con deflación. En este paquete también se incluye una rutina debida a R. P. Brent en la que se usa una combinación de interpolación lineal, una interpolación cuadrática inversa similar al método de Müller y el método de bisección. El método de Laguerre también se usa para encontrar los ceros de un polinomio real. Otra rutina para encontrar los ceros de los polinomios reales usa el método de Jenkins–Traub, que también sirve para encontrar los ceros de un polinomio complejo.

La biblioteca NAG tiene una subrutina que usa una combinación del método de bisección, la interpolación lineal y la extrapolación para aproximar un cero real de una función en un intervalo determinado. NAG también provee subrutinas para aproximar todos los ceros de un polinomio real o complejo, respectivamente. Ambas subrutinas usan un método de Laguerre modificado.

La biblioteca netlib contiene una subrutina que usa una combinación de los métodos de bisección y de secante desarrollada por T. J. Dekker para aproximar un cero real de la función en el intervalo. Requiere especificar un intervalo que contiene una raíz y regresa un intervalo con un ancho que se encuentra dentro de una tolerancia específica. Otra subrutina usa una combinación del método de bisección, la interpolación y la extrapolación para encontrar un cero real de la función en el intervalo.

Observe que a pesar de la diversidad de los métodos, los paquetes escritos de manera profesional están basados principalmente en métodos y principios que se analizan en este capítulo. Usted debería ser capaz de utilizar estos paquetes al leer los manuales adjuntos para comprender mejor los parámetros y las especificaciones de los resultados obtenidos.

Existen tres libros que consideramos clásicos para la solución de ecuaciones no lineales, los de Traub [Tr], de Ostrowski [Os] y de Householder [Ho]. Además, el libro de Brent [Bre] sirvió como base para muchos de los métodos para encontrar la raíz que se usan en la actualidad.

Las secciones Preguntas de análisis, Conceptos clave y Revisión del capítulo están disponibles en línea. Encuentre la ruta de acceso en las páginas preliminares.

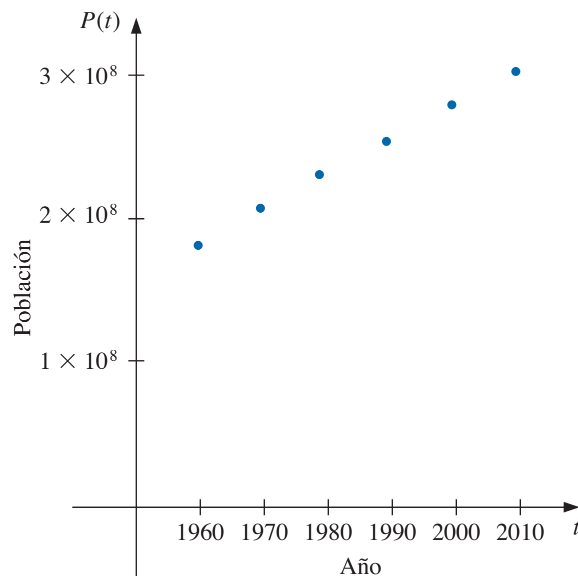
CAPÍTULO

3 Interpolación y aproximación polinomial

Introducción

Se realiza un censo de la población de Estados Unidos cada 10 años. La siguiente tabla muestra la población, en miles de personas, desde 1960 hasta 2010, y los datos también se representan en la figura.

Año	1960	1970	1980	1990	2000	2010
Población (en miles)	179 323	203 302	226 542	249 633	281 422	308 746



Al revisar estos datos, podríamos preguntar si se podrían usar para efectuar un cálculo razonable de la población, digamos, en 1975 o incluso en el año 2020. Las predicciones de este tipo pueden obtenerse por medio de una función que se ajuste a los datos proporcionados. Este proceso recibe el nombre de *interpolación* y es el tema de este capítulo. Este problema de población se considera a lo largo del capítulo y en los ejercicios 19 de la sección 3.1, 17 de la sección 3.3 y 24 de la sección 3.5.



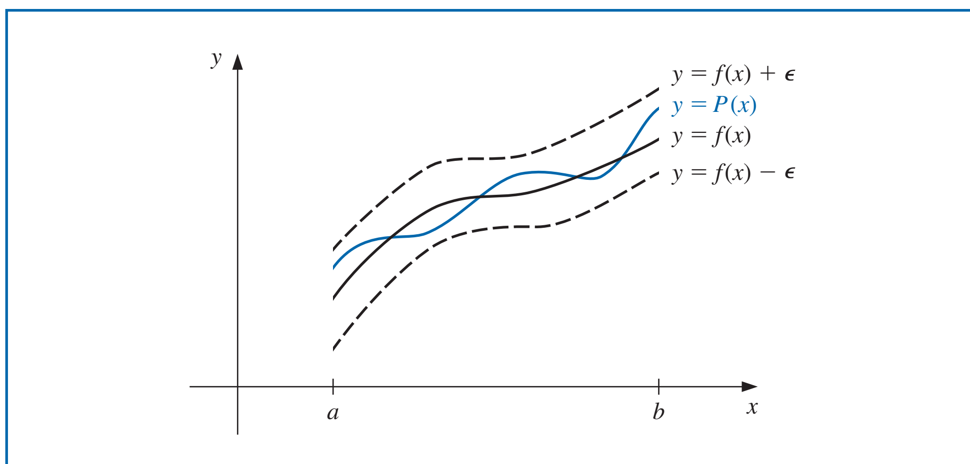
3.1 Interpolación y el polinomio de Lagrange

Una de las clases más útiles y conocidas de funciones que mapean el conjunto de números reales en sí mismo son los *polinomios algebraicos*, el conjunto de funciones de la forma

$$P_n(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0,$$

donde n es un entero positivo y $a_0, \ldots, a_n$ son constantes reales. Una razón de su importancia es que se aproximan de manera uniforme a las funciones continuas. Con esto queremos decir que dada una función, definida y continua sobre un intervalo cerrado y acotado, existe un polinomio que está tan "cerca" de la función dada como se desee. Este resultado se expresa con precisión en el teorema de aproximación de Weierstrass (consulte la figura 3.1).

Figura 3.1



Teorema 3.1 (Teorema de aproximación de Weierstrass)

Suponga que f está definida y es continua en $[a, b]$. Para cada $\epsilon > 0$, existe un polinomio $P(x)$, con la propiedad de que

$$|f(x) - P(x)| < \epsilon, \quad \text{para todas las } x \text{ en } [a, b]. \quad ■$$

A menudo, se hace referencia a Karl Weierstrass (1815–1897) como el padre del análisis moderno debido a su insistencia sobre el rigor en la demostración de resultados matemáticos. Fue fundamental para el desarrollo de pruebas de convergencia de series y para determinar formas de definir rigurosamente los números irracionales. Fue el primero en demostrar que una función podría ser continua en todas partes, pero diferenciable en ninguna parte, un resultado que escandalizó a algunos de sus contemporáneos.

La prueba de este teorema se puede encontrar en la mayoría de los textos básicos sobre análisis real (consulte, por ejemplo, [Bart], pp. 165–172).

Otra razón importante para considerar la clase de polinomios en la aproximación de funciones es que la derivada y la integral indefinida de un polinomio son fáciles de determinar y también son polinomios. Por esta razón, a menudo se usan polinomios para aproximar funciones continuas.

Los polinomios de Taylor se presentaron en la sección 1.1, donde se describieron como uno de los componentes básicos del análisis numérico. Debido a su importancia, se podría esperar que la aproximación polinomial usará estas funciones en gran medida; sin embargo, éste no es el caso. Los polinomios de Taylor concuerdan tanto como es posible con una función dada en un punto específico, pero concentran su precisión cerca de ese punto. Un buen polinomio de aproximación debe dar precisión relativa sobre un intervalo completo y, en general, los polinomios de Taylor no lo hacen. Por ejemplo, suponga que calculamos los

primeros seis polinomios de Taylor alrededor de $x_0 = 0$ para $f(x) = e^x$. Ya que las derivadas de $f(x)$ son todas e^x, que evaluadas en $x_0 = 0$ dan 1, los polinomios de Taylor son

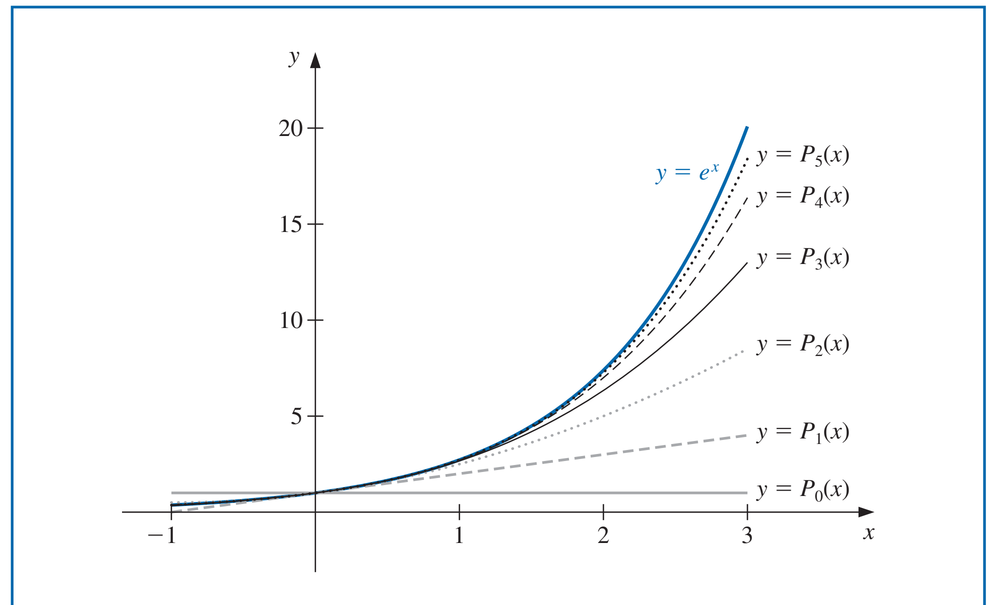
Se publicó muy poco del trabajo de Weierstrass durante su vida; no obstante, sus conferencias, en especial sobre la teoría de las funciones, influyeron de manera significativa en una generación completa de estudiantes.

$$P_0(x) = 1, \quad P_1(x) = 1 + x, \quad P_2(x) = 1 + x + \frac{x^2}{2}, \quad P_3(x) = 1 + x + \frac{x^2}{2} + \frac{x^3}{6},$$

$$P_4(x) = 1 + x + \frac{x^2}{2} + \frac{x^3}{6} + \frac{x^4}{24}, \quad \text{y} \quad P_5(x) = 1 + x + \frac{x^2}{2} + \frac{x^3}{6} + \frac{x^4}{24} + \frac{x^5}{120}.$$

Las gráficas de los polinomios se muestran en la figura 3.2 (observe que incluso para los polinomios de grado más alto, el error empeora progresivamente conforme nos alejamos de cero).

Figura 3.2



Aunque se obtienen mejores aproximaciones para $f(x) = e^x$ si se usan polinomios de Taylor, esto no es verdad para todas las funciones. Considere, como un ejemplo extremo, usar la expansión en polinomios de Taylor de diferentes grados para $f(x) = 1/x$ alrededor de $x_0 = 1$ para aproximar $f(3) = 1/3$. Puesto que

$$f(x) = x^{-1}, \quad f'(x) = -x^{-2}, \quad f''(x) = (-1)^2 2 \cdot x^{-3},$$

y, en general,

$$f^{(k)}(x) = (-1)^k k! x^{-k-1},$$

los polinomios de Taylor son

$$P_n(x) = \sum_{k=0}^{n} \frac{f^{(k)}(1)}{k!}(x-1)^k = \sum_{k=0}^{n} (-1)^k (x-1)^k.$$

Para aproximar $f(3) = 1/3$ mediante $P_n(3)$ para valores cada vez mayores de n, obtenemos los valores en la tabla 3.1 (¡un terrible fracaso!). Cuando aproximamos $f(3) = 1/3$ mediante $P_n(3)$ y para valores más grandes de n, la aproximación se vuelve cada vez más imprecisa.

Tabla 3.1

n	0	1	2	3	4	5	6	7
$P_n(3)$	1	−1	3	−5	11	−21	43	−85

Para los polinomios de Taylor, toda la información que se usa en la aproximación se concentra en el único número x_0, por lo que, en general, éstos darán aproximaciones imprecisas conforme nos alejamos de x_0. Esto limita la aproximación de polinomios de Taylor a situaciones en las que las aproximaciones sólo se necesitan en números cercanos a x_0. Para propósitos computacionales ordinarios, es más eficiente usar métodos que incluyan información en varios puntos. Consideramos esto en el resto del capítulo. El uso principal de los polinomios de Taylor en el análisis numérico no tiene propósitos de aproximación, sino la derivación de técnicas numéricas y el cálculo de errores.

Polinomios de interpolación de Lagrange

El problema de determinar un polinomio de grado uno que pasa por diferentes puntos (x_0, y_0) y (x_1, y_1) es igual al de aproximar una función f para la que $f(x_0) = y_0$ y $f(x_1) = y_1$ por medio de un polinomio de primer grado que se **interpola**, o que coincida con los valores de f en los puntos determinados. El uso de estos polinomios para aproximación dentro del intervalo determinado mediante puntos finales recibe el nombre de **interpolación**.

Defina las funciones

$$L_0(x) = \frac{x - x_1}{x_0 - x_1} \quad \text{y} \quad L_1(x) = \frac{x - x_0}{x_1 - x_0}.$$

El **polinomio de interpolación de Lagrange** lineal a través de (x_0, y_0) y (x_1, y_1) es

$$P(x) = L_0(x)f(x_0) + L_1(x)f(x_1) = \frac{x - x_1}{x_0 - x_1}f(x_0) + \frac{x - x_0}{x_1 - x_0}f(x_1).$$

Observe que

$$L_0(x_0) = 1, \quad L_0(x_1) = 0, \quad L_1(x_0) = 0, \quad \text{y} \quad L_1(x_1) = 1,$$

lo cual implica que

$$P(x_0) = 1 \cdot f(x_0) + 0 \cdot f(x_1) = f(x_0) = y_0$$

y

$$P(x_1) = 0 \cdot f(x_0) + 1 \cdot f(x_1) = f(x_1) = y_1.$$

Por lo que P es el único polinomio de grado a lo más 1 que pasa por (x_0, y_0) y (x_1, y_1).

Ejemplo 1 Determine el polinomio de interpolación de Lagrange que pasa por los puntos (2, 4) y (5, 1).

Solución en este caso, tenemos

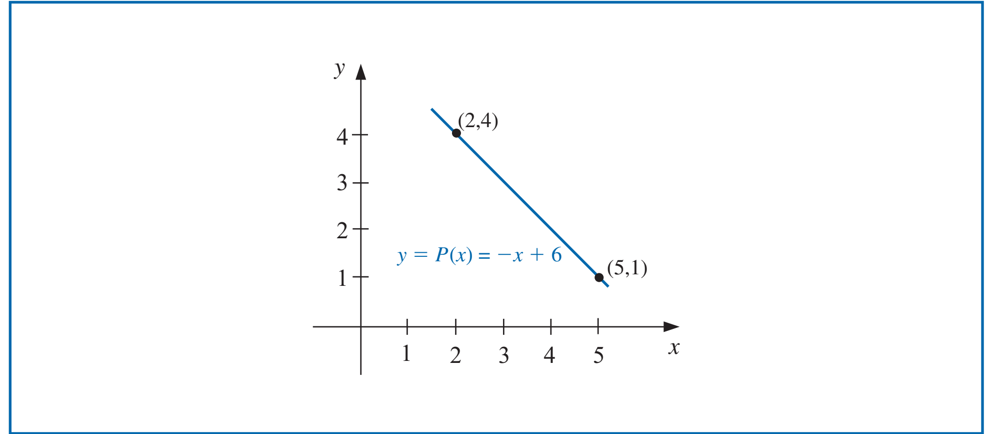
$$L_0(x) = \frac{x - 5}{2 - 5} = -\frac{1}{3}(x - 5) \quad \text{y} \quad L_1(x) = \frac{x - 2}{5 - 2} = \frac{1}{3}(x - 2),$$

por lo que

$$P(x) = -\frac{1}{3}(x - 5) \cdot 4 + \frac{1}{3}(x - 2) \cdot 1 = -\frac{4}{3}x + \frac{20}{3} + \frac{1}{3}x - \frac{2}{3} = -x + 6.$$

La gráfica de $y = P(x)$ se muestra en la figura 3.3. ■

Figura 3.3

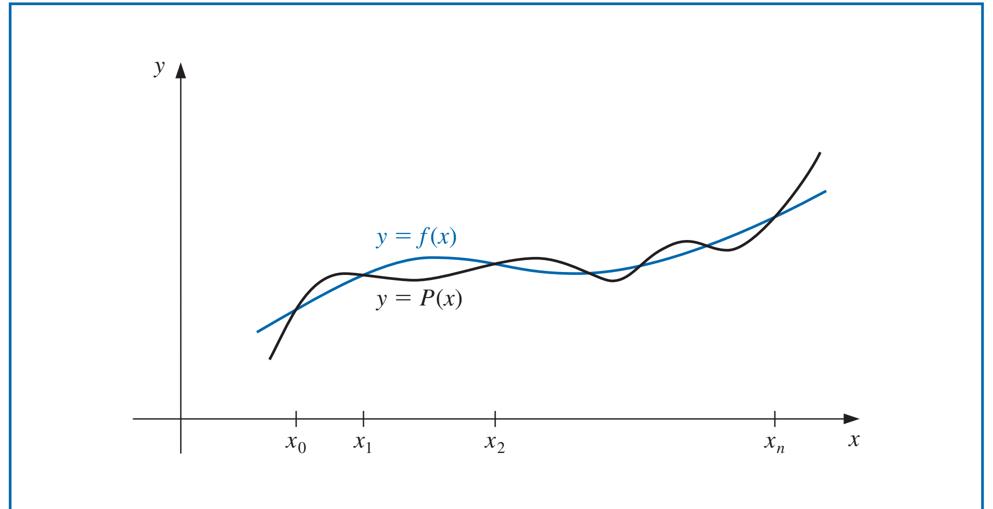


Para generalizar el concepto de interpolación lineal, considere la construcción de un polinomio de grado n que pasa a través de $n + 1$ puntos

$$(x_0, f(x_0)),\ (x_1, f(x_1)), \ldots, (x_n, f(x_n)).$$

(Véase la figura 3.4.)

Figura 3.4



En este caso, primero construimos, para cada $k = 0, 1, \ldots, n$, una función $L_{n,k}(x)$ con la propiedad de que $L_{n,k}(x_i) = 0$ cuando $i \neq k$ y $L_{n,k}(x_k) = 1$. Para satisfacer $L_{n,k}(x_i) = 0$ para cada $i \neq k$ se requiere que el numerador de $L_{n,k}(x)$ contenga el término

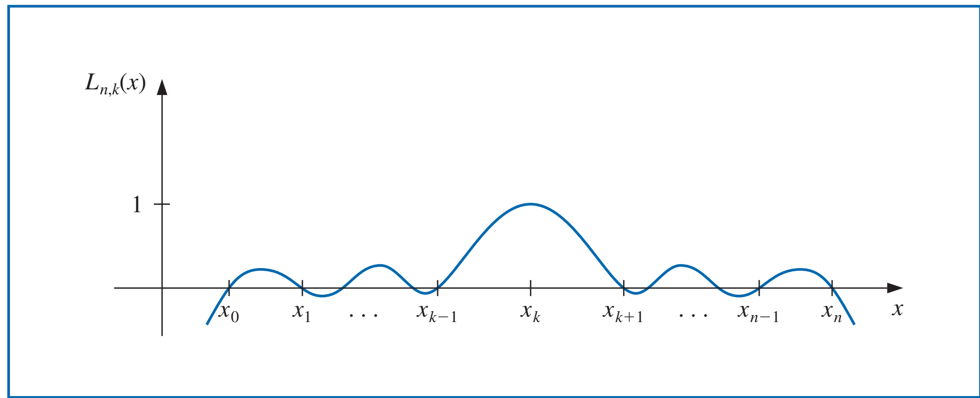
$$(x - x_0)(x - x_1) \cdots (x - x_{k-1})(x - x_{k+1}) \cdots (x - x_n).$$

Para satisfacer $L_{n,k}(x_k) = 1$, el denominador de $L_{n,k}(x)$ debe ser el mismo término, pero evaluado en $x = x_k$. Por lo tanto,

$$L_{n,k}(x) = \frac{(x - x_0) \cdots (x - x_{k-1})(x - x_{k+1}) \cdots (x - x_n)}{(x_k - x_0) \cdots (x_k - x_{k-1})(x_k - x_{k+1}) \cdots (x_k - x_n)}.$$

Un bosquejo de la gráfica de una $L_{n,k}$ (cuando n es par) se muestra en la figura 3.5.

Figura 3.5



El polinomio de interpolación se describe fácilmente una vez que se conoce la forma $L_{n,k}$. Este polinomio, llamado **enésimo polinomio de interpolación de Lagrange,** se define en el siguiente teorema.

Teorema 3.2 Si $x_0, x_1, \ldots, x_n$ son $n+1$ números distintos y f es una función cuyos valores están determinados en estos números, entonces existe un único polinomio $P(x)$ de grado a lo sumo n con

$$f(x_k) = P(x_k), \quad \text{para cada } k = 0, 1, \ldots, n.$$

Este polinomio está determinado por

$$P(x) = f(x_0)L_{n,0}(x) + \cdots + f(x_n)L_{n,n}(x) = \sum_{k=0}^{n} f(x_k)L_{n,k}(x), \tag{3.1}$$

donde, para cada $k = 0, 1, \ldots, n$,

$$L_{n,k}(x) = \frac{(x-x_0)(x-x_1)\cdots(x-x_{k-1})(x-x_{k+1})\cdots(x-x_n)}{(x_k-x_0)(x_k-x_1)\cdots(x_k-x_{k-1})(x_k-x_{k+1})\cdots(x_k-x_n)} \tag{3.2}$$

$$= \prod_{\substack{i=0 \\ i\neq k}}^{n} \frac{(x-x_i)}{(x_k-x_i)}.$$

■

La fórmula de interpolación nombrada por Joseph Louis Lagrange (1736–1813) probablemente era conocida por Newton alrededor de 1675, pero al parecer fue publicada por primera vez en 1779 por Edward Waring (1736–1798). Lagrange escribió mucho sobre el tema de interpolación y su trabajo tuvo una influencia significativa sobre los matemáticos posteriores. Él publicó este resultado en 1795.

El símbolo $\prod$ se usa para escribir productos de manera compacta y es similar al símbolo $\sum$, que se utiliza para escribir sumas. Por ejemplo $\prod_{i=0}^{3} a_i = a_1 * a_2 * a_3$.

Escribiremos $L_{n,k}(x)$ simplemente como $L_k(x)$ cuando no haya confusión en cuanto a su grado.

Ejemplo 2

a) Use los números (llamados *nodos*)

$x_0 = 2$, $x_1 = 2.75$ y $x_2 = 4$ para encontrar el polinomio de interpolación de Lagrange de segundo grado para $f(x) = 1/x$.

b) Use este polinomio para aproximar $f(3) = 1/3$.

Solución **a)** Primero determinamos los coeficientes polinómicos $L_0(x)$, $L_1(x)$ y $L_2(x)$. En forma anidada, estos son

$$L_0(x) = \frac{(x-2.75)(x-4)}{(2-2.75)(2-4)} = \frac{2}{3}(x-2.75)(x-4),$$

$$L_1(x) = \frac{(x-2)(x-4)}{(2.75-2)(2.75-4)} = -\frac{16}{15}(x-2)(x-4),$$

y

$$L_2(x) = \frac{(x-2)(x-2.75)}{(4-2)(4-2.75)} = \frac{2}{5}(x-2)(x-2.75).$$

Además, $f(x_0) = f(2) = 1/2$, $f(x_1) = f(2.75) = 4/11$, y $f(x_2) = f(4) = 1/4$, por lo que

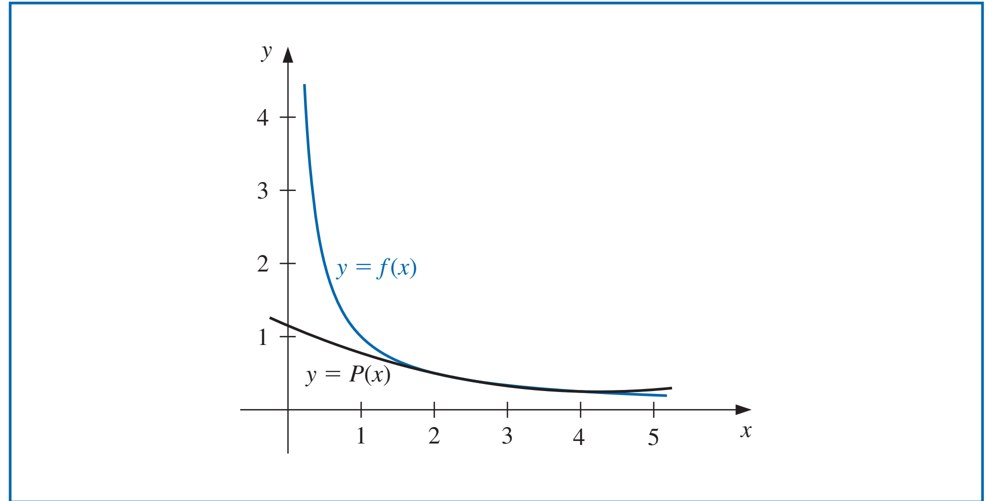
$$\begin{aligned}
P(x) &= \sum_{k=0}^{2} f(x_k)L_k(x) \\
&= \frac{1}{3}(x-2.75)(x-4) - \frac{64}{165}(x-2)(x-4) + \frac{1}{10}(x-2)(x-2.75) \\
&= \frac{1}{22}x^2 - \frac{35}{88}x + \frac{49}{44}.
\end{aligned}$$

b) Una aproximación para $f(3) = 1/3$ (véase la figura 3.6) es

$$f(3) \approx P(3) = \frac{9}{22} - \frac{105}{88} + \frac{49}{44} = \frac{29}{88} \approx 0.32955.$$

Recuerde que en la sección de apertura de este capítulo (consulte la tabla 3.1), encontramos que ninguna expansión en polinomios de Taylor alrededor de $x_0 = 1$ se puede usar para aproximar razonablemente $f(x) = 1/x$ en $x = 3$. ■

Figura 3.6



El siguiente paso es calcular un residuo o cota para el error involucrado en la aproximación de una función mediante un polinomio de interpolación.

Teorema 3.3 Suponga $x_0, x_1, \ldots, x_n$ son números distintos en el intervalo $[a, b]$ y $f \in C^{n+1}[a, b]$. Entonces, para cada x en $[a, b]$, existe un número $\xi(x)$ (generalmente no conocido) entre mín $\{x_0, x_1, \ldots, x_n\}$ y máx$\{x_0, x_1, \ldots, x_n\}$ y, por lo tanto, en (a, b), con

Existen otras formas de expresar el término de error para el polinomio de Lagrange, pero ésta puede ser la forma más útil y la que concuerda más estrechamente con la forma de error del polinomio estándar de Taylor.

$$f(x) = P(x) + \frac{f^{(n+1)}(\xi(x))}{(n+1)!}(x-x_0)(x-x_1)\cdots(x-x_n), \tag{3.3}$$

donde $P(x)$ es el polinomio de interpolación determinado en la ecuación (3.1).

Demostración Primero observe que si $x = x_k$ para cualquier $k = 0, 1, \ldots, n$, entonces $f(x_k) = P(x_k)$ y al elegir $\xi(x_k)$de manera arbitraria en (a, b) se obtiene la ecuación (3.3).

Si $x \neq x_k$, para todas las $k = 0, 1, \ldots, n$ defina la función g para t en $[a, b]$ mediante

$$\begin{aligned} g(t) &= f(t) - P(t) - [f(x) - P(x)]\frac{(t-x_0)(t-x_1)\cdots(t-x_n)}{(x-x_0)(x-x_1)\cdots(x-x_n)} \\ &= f(t) - P(t) - [f(x) - P(x)]\prod_{i=0}^{n}\frac{(t-x_i)}{(x-x_i)}. \end{aligned}$$

Puesto que $f \in C^{n+1}[a, b]$, y $P \in C^{\infty}[a, b]$, se sigue que $g \in C^{n+1}[a, b]$. Para $t = x_k$, tenemos

$$g(x_k) = f(x_k) - P(x_k) - [f(x) - P(x)]\prod_{i=0}^{n}\frac{(x_k-x_i)}{(x-x_i)} = 0 - [f(x) - P(x)] \cdot 0 = 0.$$

Además,

$$g(x) = f(x) - P(x) - [f(x) - P(x)]\prod_{i=0}^{n}\frac{(x-x_i)}{(x-x_i)} = f(x) - P(x) - [f(x) - P(x)] = 0.$$

Por lo tanto, $g \in C^{n+1}[a, b]$, y g se anula en los $n + 2$ números distintos $x, x_0, x_1, \ldots, x_n$. Por el teorema generalizado de Rolle 1.10, existe un número ξ en (a, b) para el que $g^{(n+1)}(\xi) = 0$.Por lo que,

$$0 = g^{(n+1)}(\xi) = f^{(n+1)}(\xi) - P^{(n+1)}(\xi) - [f(x) - P(x)]\frac{d^{n+1}}{dt^{n+1}}\left[\prod_{i=0}^{n}\frac{(t-x_i)}{(x-x_i)}\right]_{t=\xi}. \quad (3.4)$$

Sin embargo, $P(x)$ es un polinomio de grado a lo sumo n, por lo que la derivada $(n + 1)$, $P^{(n+1)}(x)$, es cero. Además $\prod_{i=0}^{n}[(t - x_i)/(x - x_i)]$ es un polinomio de grado $(n + 1)$, por lo que

$$\prod_{i=0}^{n}\frac{(t-x_i)}{(x-x_i)} = \left[\frac{1}{\prod_{i=0}^{n}(x-x_i)}\right]t^{n+1} + (\text{términos de menor grado en } t),$$

y

$$\frac{d^{n+1}}{dt^{n+1}}\prod_{i=0}^{n}\frac{(t-x_i)}{(x-x_i)} = \frac{(n+1)!}{\prod_{i=0}^{n}(x-x_i)}.$$

Ahora, la ecuación (3.4) se convierte en

$$0 = f^{(n+1)}(\xi) - 0 - [f(x) - P(x)]\frac{(n+1)!}{\prod_{i=0}^{n}(x-x_i)},$$

y, después de resolver $f(x)$, tenemos

$$f(x) = P(x) + \frac{f^{(n+1)}(\xi)}{(n+1)!}\prod_{i=0}^{n}(x-x_i). \qquad ■$$

La fórmula de error en el teorema 3.3 es un resultado teórico importante porque los polinomios de Lagrange se usan ampliamente para deducir la diferenciación numérica y los métodos de integración. Las cotas de error para estas técnicas se obtienen a partir de la fórmula del error de Lagrange.

Observe que la forma del error para el polinomio de Lagrange es bastante similar a la del polinomio de Taylor. El enésimo polinomio de Taylor alrededor de x_0 concentra toda la información conocida en x_0 y tiene un término de error de la forma

$$\frac{f^{(n+1)}(\xi(x))}{(n+1)!}(x - x_0)^{n+1}.$$

El polinomio de Lagrange de grado n utiliza información en los distintos números x_0, x_1, . . . , x_n y, en lugar de $(x - x_0)^n$ su fórmula de error utiliza el producto de los $n + 1$ términos $(x - x_0), (x - x_1), \ldots, (x - x_n)$:

$$\frac{f^{(n+1)}(\xi(x))}{(n+1)!}(x - x_0)(x - x_1)\cdots(x - x_n).$$

Ejemplo 3 En el ejemplo 2 encontramos el segundo polinomio de Lagrange para $f(x) = 1/x$ en $[2, 4]$ usando los nodos $x_0 = 2$, $x_1 = 2.75$ y $x_2 = 4$. Determine la forma del error para este polinomio y el error máximo cuando el polinomio se usa para aproximar $f(x)$ para $x \in [2, 4]$.

Solución Como $f(x) = x^{-1}$, tenemos

$$f'(x) = -x^{-2}, \quad f''(x) = 2x^{-3}, \quad \text{y} \quad f'''(x) = -6x^{-4}.$$

En consecuencia, el segundo polinomio de Lagrange tiene el error de la forma

$$\frac{f'''(\xi(x))}{3!}(x-x_0)(x-x_1)(x-x_2) = -\,(\xi(x))^{-4}(x-2)(x-2.75)(x-4), \quad \text{para } \xi(x) \text{ en } (2, 4).$$

El valor máximo de $(\xi(x))^{-4}$ en el intervalo es $2^{-4} = 1/16$. Ahora necesitamos determinar el valor máximo en este intervalo del valor absoluto del polinomio

$$g(x) = (x-2)(x-2.75)(x-4) = x^3 - \frac{35}{4}x^2 + \frac{49}{2}x - 22.$$

Como

$$D_x\left(x^3 - \frac{35}{4}x^2 + \frac{49}{2}x - 22\right) = 3x^2 - \frac{35}{2}x + \frac{49}{2} = \frac{1}{2}(3x-7)(2x-7),$$

los puntos críticos se presentan en

$$x = \frac{7}{3}, \text{ con } g\left(\frac{7}{3}\right) = \frac{25}{108}, \qquad \text{y} \qquad x = \frac{7}{2}, \text{ con } g\left(\frac{7}{2}\right) = -\frac{9}{16}.$$

Por lo tanto, el error máximo es

$$\frac{f'''(\xi(x))}{3!}|(x-x_0)(x-x_1)(x-x_2)| \leq \frac{1}{16}\left|-\frac{9}{16}\right| = \frac{9}{256} \approx 0.03515625.$$ ■

El siguiente ejemplo ilustra cómo se puede usar la fórmula del error para preparar una tabla de datos que garantizará un error de interpolación dentro de una cota establecida.

Ejemplo 4 Suponga que se va a preparar una tabla para la función $f(x) = e^x$, para x en $[0, 1]$. Imagine que el número de lugares decimales proporcionado por entrada es $d \geq 8$ y que h, el tamaño del paso es la diferencia entre valores adyacentes x. ¿Qué tamaño de paso h garantizará que la interpolación lineal proporcione un error absoluto a lo máximo de 10^{-6} para todas las x en $[0, 1]$?

Solución Sean $x_0, x_1, \ldots$ los números en los que se evalúa f y x está en $[0, 1]$ y suponga que j satisface $x_j \leq x \leq x_{j+1}$. La ecuación (3.3) implica que el error en la interpolación lineal es

$$|f(x) - P(x)| = \left|\frac{f^{(2)}(\xi)}{2!}(x - x_j)(x - x_{j+1})\right| = \frac{|f^{(2)}(\xi)|}{2}|(x - x_j)||(x - x_{j+1})|.$$

Como el tamaño del paso es h, entonces $x_j = jh$, $x_{j+1} = (j+1)h$, y

$$|f(x) - P(x)| \leq \frac{|f^{(2)}(\xi)|}{2!}|(x - jh)(x - (j+1)h)|.$$

Por lo tanto,

$$|f(x) - P(x)| \leq \frac{\text{máx}_{\xi\in[0,1]} e^{\xi}}{2} \max_{x_j \leq x \leq x_{j+1}} |(x - jh)(x - (j+1)h)|$$
$$\leq \frac{e}{2} \max_{x_j \leq x \leq x_{j+1}} |(x - jh)(x - (j+1)h)|.$$

Considere la función $g(x) = (x - jh)(x - (j+1)h)$, para $jh \leq x \leq (j+1)h$. Luego

$$g'(x) = (x - (j+1)h) + (x - jh) = 2\left(x - jh - \frac{h}{2}\right),$$

el único punto crítico para g se encuentra en $x = jh + h/2$, con $g(jh + h/2) = (h/2)^2 = h^2/4$.

Puesto que $g(jh) = 0$ y $g((j+1)h) = 0$, el valor máximo de $|g'(x)|$ en $[jh, (j+1)h]$ se debe presentar en el punto crítico, lo cual implica que (véase el ejercicio 21)

$$|f(x) - P(x)| \leq \frac{e}{2} \max_{x_j \leq x \leq x_{j+1}} |g(x)| \leq \frac{e}{2} \cdot \frac{h^2}{4} = \frac{eh^2}{8}.$$

Por consiguiente, para garantizar que el error en la interpolación lineal está acotado por 10^{-6}, es suficiente elegir h de tal forma que

$$\frac{eh^2}{8} \leq 10^{-6}. \quad \text{Esto implica que} \quad h < 1.72 \times 10^{-3}.$$

Puesto que $n = (1 - 0)/h$ debe ser un entero, una selección razonable para el tamaño del paso es $h = 0.001$. ■

La sección Conjunto de ejercicios 3.1 está disponible en línea. Encuentre la ruta de acceso en las páginas preliminares.

3.2 Aproximación de datos y método de Neville

En la sección anterior encontramos una representación explícita para los polinomios de Lagrange y su error cuando se aproxima una función sobre un intervalo. El uso frecuente de estos polinomios implica la interpolación de datos tabulados. En este caso, una representación explícita del polinomio podría no ser necesaria, sólo los valores del polinomio en puntos específicos. En esta situación, sería posible que la función subyacente a los datos no se conozca, por lo que la forma explícita del error no se puede usar. Ahora, ilustraremos una aplicación práctica de interpolación en dicha situación.

Ilustración La tabla 3.2 lista los valores de una función f en diferentes puntos. Las aproximaciones para $f(1.5)$ obtenidas con distintos polinomios de Lagrange que usan estos datos se comparará para probar y determinar la precisión de la aproximación.

Tabla 3.2

x	$f(x)$
1.0	0.7651977
1.3	0.6200860
1.6	0.4554022
1.9	0.2818186
2.2	0.1103623

El polinomio lineal más apropiado usa $x_0 = 1.3$ y $x_1 = 1.6$ porque 1.5 se encuentra entre 1.3 y 1.6. El valor del polinomio de interpolación en 1.5 es

$$P_1(1.5) = \frac{(1.5 - 1.6)}{(1.3 - 1.6)} f(1.3) + \frac{(1.5 - 1.3)}{(1.6 - 1.3)} f(1.6)$$
$$= \frac{(1.5 - 1.6)}{(1.3 - 1.6)}(0.6200860) + \frac{(1.5 - 1.3)}{(1.6 - 1.3)}(0.4554022) = 0.5102968.$$

Es posible usar razonablemente dos polinomios de grado dos, uno con $x_0 = 1.3$, $x_1 = 1.6$ y $x_2 = 1.9$, lo cual nos da

$$P_2(1.5) = \frac{(1.5-1.6)(1.5-1.9)}{(1.3-1.6)(1.3-1.9)}(0.6200860) + \frac{(1.5-1.3)(1.5-1.9)}{(1.6-1.3)(1.6-1.9)}(0.4554022)$$
$$+ \frac{(1.5-1.3)(1.5-1.6)}{(1.9-1.3)(1.9-1.6)}(0.2818186) = 0.5112857,$$

y uno con $x_0 = 1.0$, $x_1 = 1.3$ y $x_2 = 1.6$, lo cual nos da $\hat{P}_2(1.5) = 0.5124715$.

En el caso de tercer grado, también hay dos opciones razonables para el polinomio, una con $x_0 = 1.3$, $x_1 = 1.6$, $x_2 = 1.9$ y $x_3 = 2.2$, lo cual nos da $P_3(1.5) = 0.5118302$. La segunda aproximación de tercer grado se obtiene con $x_0 =1.0$, $x_1 = 1.3$, $x_2 = 1.6$ y $x_3 = 1.9$, lo cual nos da $\hat{P}_3(1.5) = 0.5118127$.

El polinomio de Lagrange de cuarto grado usa todas las entradas en la tabla. Con $x_0 = 1.0$, $x_1 = 1.3$, $x_2 = 1.6$, $x_3 = 1.9$ y $x_4 = 2.2$, la aproximación es $P_4(1.5) = 0.5118200$.

Puesto que $P_3(1.5)$, $\hat{P}_3(1.5)$ y $P_4(1.5)$ concuerdan con una exactitud de 2×10^{-5} unidades, esperamos este grado de precisión para estas aproximaciones. También esperamos que $P_4(1.5)$ sea la aproximación más precisa ya que usa la mayor parte de los datos proporcionados.

La función que estamos aproximando es, en realidad, la función de Bessel de primera clase de orden cero, cuyo valor en 1.5 se conoce como 0.5118277. Por lo tanto, las verdaderas precisiones de las aproximaciones son las siguientes:

$$|P_1(1.5) - f(1.5)| \approx 1.53 \times 10^{-3},$$
$$|P_2(1.5) - f(1.5)| \approx 5.42 \times 10^{-4},$$
$$|\hat{P}_2(1.5) - f(1.5)| \approx 6.44 \times 10^{-4},$$
$$|P_3(1.5) - f(1.5)| \approx 2.5 \times 10^{-6},$$
$$|\hat{P}_3(1.5) - f(1.5)| \approx 1.50 \times 10^{-5},$$
$$|P_4(1.5) - f(1.5)| \approx 7.7 \times 10^{-6}.$$

Aunque $P_3(1.5)$ es la aproximación más precisa, si no conocemos el valor real de $f(1.5)$, aceptaríamos $P_4(1.5)$ como la mejor aproximación ya que incluye la mayor cantidad de datos sobre la función. El término del error de Lagrange derivado del teorema 3.3 no se puede aplicar aquí porque no conocemos la cuarta derivada de f. Por desgracia, este casi siempre es el caso. ■

Método de Neville

Una dificultad práctica con la interpolación de Lagrange es que el término del error es difícil de aplicar, por lo que el grado del polinomio que se necesita para la precisión deseada en general se desconoce hasta que se realizan los cálculos. Una práctica común es calcular los resultados dados a partir de diferentes polinomios hasta que se obtiene el acuerdo apropiado, como se hizo en la ilustración anterior. Sin embargo, el trabajo efectuado al calcular la aproximación con el segundo polinomio no disminuye el trabajo necesario para calcular la tercera aproximación, ni la cuarta aproximación es fácil de obtener una vez que se conoce la tercera aproximación y así sucesivamente. Ahora, derivaremos estos polinomios de aproximación de una manera que use los cálculos previos para una mayor ventaja.

Definición 3.4 Sea f una función definida en $x_0, x_1, x_2, \ldots, x_n$ y suponga que $m_1, m_2, \ldots, m_k$ son k enteros diferentes, con $0 \leq m_i \leq n$ para cada i. El polinomio de Lagrange que concuerda con $f(x)$ en los puntos k $x_{m_1}, x_{m_2}, \ldots, x_{m_k}$ se denota $P_{m_1, m_2, \ldots, m_k}(x)$. ■

Ejemplo 1 Suponga que $x_0 = 1$, $x_1 = 2$, $x_2 = 3$, $x_3 = 4$, $x_4 = 6$ y $f(x) = e^x$. Determine el polinomio de interpolación que se denota $P_{1,2,4}(x)$ y use este polinomio para aproximar $f(5)$.

Solución Éste es el polinomio de Lagrange que concuerda con $f(x)$ en $x_1 = 2$, $x_2 = 3$ y $x_4 = 6$. Por lo tanto,

$$P_{1,2,4}(x) = \frac{(x-3)(x-6)}{(2-3)(2-6)}e^2 + \frac{(x-2)(x-6)}{(3-2)(3-6)}e^3 + \frac{(x-2)(x-3)}{(6-2)(6-3)}e^6.$$

por lo que,

$$f(5) \approx P(5) = \frac{(5-3)(5-6)}{(2-3)(2-6)}e^2 + \frac{(5-2)(5-6)}{(3-2)(3-6)}e^3 + \frac{(5-2)(5-3)}{(6-2)(6-3)}e^6$$
$$= -\frac{1}{2}e^2 + e^3 + \frac{1}{2}e^6 \approx 218.105.$$

■

El siguiente resultado describe un método para generar de forma recursiva las aproximaciones del polinomio de Lagrange.

Teorema 3.5 Sea f definida en $x_0, x_1, \ldots, x_k$ y sean x_j y x_i dos números distintos en este conjunto. Entonces

$$P(x) = \frac{(x-x_j)P_{0,1,\ldots,j-1,j+1,\ldots,k}(x) - (x-x_i)P_{0,1,\ldots,i-1,i+1,\ldots,k}(x)}{(x_i - x_j)}$$

es el k-ésimo polinomio de Lagrange que interpola f en los puntos $k+1$ $x_0, x_1, \ldots, x_k$.

Demostración Para la facilidad de la notación, sea $Q \equiv P_{0,1,\ldots,i-1,i+1,\ldots,k}$ y $\hat{Q} \equiv P_{0,1,\ldots,j-1,j+1,\ldots,k}$. Puesto que $Q(x)$ y $\hat{Q}(x)$ son polinomios de grado $k-1$ o menos, $P(x)$ es de grado máximo k.

Primero, observe que $\hat{Q}(x_i) = f(x_i)$ implica que

$$P(x_i) = \frac{(x_i - x_j)\hat{Q}(x_i) - (x_i - x_i)Q(x_i)}{x_i - x_j} = \frac{(x_i - x_j)}{(x_i - x_j)}f(x_i) = f(x_i).$$

Similarmente, como $Q(x_j) = f(x_j)$, tenemos que $P(x_j) = f(x_j)$.

Además, si $0 \le r \le k$ y r no es i ni j, entonces $Q(x_r) = \hat{Q}(x_r) = f(x_r)$. Por lo tanto,

$$P(x_r) = \frac{(x_r - x_j)\hat{Q}(x_r) - (x_r - x_i)Q(x_r)}{x_i - x_j} = \frac{(x_i - x_j)}{(x_i - x_j)}f(x_r) = f(x_r).$$

Pero, por definición $P_{0,1,\ldots,k}(x)$ es el único polinomio de grado máximo k que concuerda con f en $x_0, x_1, \ldots, x_k$. Por lo tanto $P \equiv P_{0,1,\ldots,k}$. ■

El teorema 3.5 implica que los polinomios de interpolación pueden generarse de manera recursiva. Por ejemplo, tenemos

$$P_{0,1} = \frac{1}{x_1 - x_0}[(x-x_0)P_1 + (x-x_1)P_0], \quad P_{1,2} = \frac{1}{x_2 - x_1}[(x-x_1)P_2 + (x-x_2)P_1],$$
$$P_{0,1,2} = \frac{1}{x_2 - x_0}[(x-x_0)P_{1,2} + (x-x_2)P_{0,1}],$$

y así sucesivamente. Estos se generan de la manera que se muestra en la tabla 3.3, donde cada fila se completa antes de que las filas sucesivas comiencen.

Tabla 3.3

x_0	P_0				
x_1	P_1	$P_{0,1}$			
x_2	P_2	$P_{1,2}$	$P_{0,1,2}$		
x_3	P_3	$P_{2,3}$	$P_{1,2,3}$	$P_{0,1,2,3}$	
x_4	P_4	$P_{3,4}$	$P_{2,3,4}$	$P_{1,2,3,4}$	$P_{0,1,2,3,4}$

El procedimiento que usa el resultado del teorema 3.5 para generar recursivamente las aproximaciones de polinomios de interpolación recibe el nombre de **método de Neville**. La notación P que se usa en la tabla 3.3 es pesada debido al número de subíndices que se utilizan para representar las entradas. Observe, sin embargo, que mientras se construye un arreglo, sólo se necesitan dos subíndices. El procedimiento hacia abajo en la tabla corresponde al uso consecutivo de los puntos x_i con una i más grande, y el procedimiento hacia la derecha corresponde al incremento del grado del polinomio de interpolación. Puesto que los puntos aparecen de manera consecutiva en cada entrada, necesitamos describir sólo un punto de inicio y el número de puntos adicionales que se usan en la construcción de la aproximación.

Eric Harold Neville (1889–1961) aportó esta modificación de la fórmula de Lagrange en un artículo publicado en 1932. [N]

Para evitar los múltiples índices, dejamos que $Q_{i,j}(x)$ para $0 \le j \le i$, denote el polinomio de interpolación de grado j en los números $(j + 1)$ $x_{i-j}, x_{i-j+1}, \ldots, x_{i-1}, x_i$; es decir

$$Q_{i,j} = P_{i-j,i-j+1,\ldots,i-1,i}.$$

Usando esta notación obtenemos el arreglo de notación Q en la tabla 3.4.

Tabla 3.4

x_0	$P_0 = Q_{0,0}$				
x_1	$P_1 = Q_{1,0}$	$P_{0,1} = Q_{1,1}$			
x_2	$P_2 = Q_{2,0}$	$P_{1,2} = Q_{2,1}$	$P_{0,1,2} = Q_{2,2}$		
x_3	$P_3 = Q_{3,0}$	$P_{2,3} = Q_{3,1}$	$P_{1,2,3} = Q_{3,2}$	$P_{0,1,2,3} = Q_{3,3}$	
x_4	$P_4 = Q_{4,0}$	$P_{3,4} = Q_{4,1}$	$P_{2,3,4} = Q_{4,2}$	$P_{1,2,3,4} = Q_{4,3}$	$P_{0,1,2,3,4} = Q_{4,4}$

Ejemplo 2 Los valores de diferentes polinomios de interpolación en $x = 1.5$ se obtuvieron en la ilustración al inicio de esta sección usando los datos que se muestran en la tabla 3.5. Aplique el método de Neville a los datos mediante la construcción de una tabla recursiva de la forma que se observa en la tabla 3.4.

Tabla 3.5

x	$f(x)$
1.0	0.7651977
1.3	0.6200860
1.6	0.4554022
1.9	0.2818186
2.2	0.1103623

Solución Sea $x_0 = 1.0$, $x_1 = 1.3$, $x_2 = 1.6$, $x_3 = 1.9$ y $x_4 = 2.2$, entonces $Q_{0,0} = f(1.0)$, $Q_{1,0} = f(1.3)$, $Q_{2,0} = f(1.6)$, $Q_{3,0} = f(1.9)$ y $Q_{4,0} = f(2.2)$. Estos son los cinco polinomios de grado cero (constantes) que aproximan $f(1.5)$ y son iguales a los datos que se proporcionan en la tabla 3.5.

Al calcular la aproximación de primer grado $Q_{1,1}(1.5)$ obtenemos

$$\begin{aligned} Q_{1,1}(1.5) &= \frac{(x - x_0)Q_{1,0} - (x - x_1)Q_{0,0}}{x_1 - x_0} \\ &= \frac{(1.5 - 1.0)Q_{1,0} - (1.5 - 1.3)Q_{0,0}}{1.3 - 1.0} \\ &= \frac{0.5(0.6200860) - 0.2(0.7651977)}{0.3} = 0.5233449. \end{aligned}$$

De igual forma,

$$Q_{2,1}(1.5) = \frac{(1.5 - 1.3)(0.4554022) - (1.5 - 1.6)(0.6200860)}{1.6 - 1.3} = 0.5102968,$$

$$Q_{3,1}(1.5) = 0.5132634, \quad \text{y} \quad Q_{4,1}(1.5) = 0.5104270.$$

Se espera que la mejor aproximación lineal sea $Q_{2,1}$ porque 1.5 se encuentra entre $x_1 = 1.3$ y $x_2 = 1.6$.

De manera similar, las aproximaciones usando polinomios de grado superior están dadas por

$$Q_{2,2}(1.5) = \frac{(1.5 - 1.0)(0.5102968) - (1.5 - 1.6)(0.5233449)}{1.6 - 1.0} = 0.5124715,$$

$$Q_{3,2}(1.5) = 0.5112857, \quad \text{y} \quad Q_{4,2}(1.5) = 0.5137361.$$

Las aproximaciones de grado superior se generan de una manera similar y se muestran en la tabla 3.6. ■

Tabla 3.6

1.0	0.7651977				
1.3	0.6200860	0.5233449			
1.6	0.4554022	0.5102968	0.5124715		
1.9	0.2818186	0.5132634	0.5112857	0.5118127	
2.2	0.1103623	0.5104270	0.5137361	0.5118302	0.5118200

Si la última aproximación $Q_{4,4}$ no fue suficientemente precisa, sería posible seleccionar otro nodo x_5 y añadir otra fila a la tabla:

$$x_5 \quad Q_{5,0} \quad Q_{5,1} \quad Q_{5,2} \quad Q_{5,3} \quad Q_{5,4} \quad Q_{5,5}.$$

Entonces $Q_{4,4}$, $Q_{5,4}$ y $Q_{5,5}$ podrían compararse para determinar la precisión posterior.

La función en el ejemplo 2 es la función de Bessel de primera clase de orden cero, cuyo valor en 2.5 es -0.0483838, y la siguiente fila de aproximaciones para $f(1.5)$ es

$$2.5 \quad -0.0483838 \quad 0.4807699 \quad 0.5301984 \quad 0.5119070 \quad 0.5118430 \quad 0.5118277.$$

La última nueva entrada, 0.5118277, es correcta para siete lugares decimales.

Ejemplo 3 La tabla 3.7 lista los valores de $f(x) = \ln x$ precisos para los lugares dados. Use el método de Neville y la aritmética de redondeo de cuatro dígitos para aproximar $f(2.1) = \ln 2.1$ al completar la tabla de Neville.

Tabla 3.7

i	x_i	$\ln x_i$
0	2.0	0.6931
1	2.2	0.7885
2	2.3	0.8329

Solución Puesto que $x - x_0 = 0.1$, $x - x_1 = -0.1$ y $x - x_2 = -0.2$, tenemos $Q_{0,0} = 0.6931$, $Q_{1,0} = 0.7885$ y $Q_{2,0} = 0.8329$,

$$Q_{1,1} = \frac{1}{0.2}\left[(0.1)0.7885 - (-0.1)0.6931\right] = \frac{0.1482}{0.2} = 0.7410$$

y

$$Q_{2,1} = \frac{1}{0.1}\left[(-0.1)0.8329 - (-0.2)0.7885\right] = \frac{0.07441}{0.1} = 0.7441.$$

La aproximación final que podemos obtener a partir de estos datos es

$$Q_{2,1} = \frac{1}{0.3}\left[(0.1)0.7441 - (-0.2)0.7410\right] = \frac{0.2276}{0.3} = 0.7420.$$

Estos valores se muestran en la tabla 3.8. ■

Tabla 3.8

i	x_i	$x - x_i$	Q_{i0}	Q_{i1}	Q_{i2}
0	2.0	0.1	0.6931		
1	2.2	−0.1	0.7885	0.7410	
2	2.3	−0.2	0.8329	0.7441	0.7420

En el ejemplo anterior, tenemos $f(2.1) = \ln 2.1 = 0.7419$ para cuatro lugares decimales, por lo que el error absoluto es

$$|f(2.1) - P_2(2.1)| = |0.7419 - 0.7420| = 10^{-4}.$$

Sin embargo, $f'(x) = 1/x$, $f''(x) = -1/x^2$, y $f'''(x) = 2/x^3$, por lo que la fórmula de error de Lagrange (3.3) en el teorema 3.3 nos da la cota del de error

$$\begin{aligned} |f(2.1) - P_2(2.1)| &= \left| \frac{f'''(\xi(2.1))}{3!}(x - x_0)(x - x_1)(x - x_2) \right| \\ &= \left| \frac{1}{3\,(\xi(2.1))^3}(0.1)(-0.1)(-0.2) \right| \leq \frac{0.002}{3(2)^3} = 8.\overline{3} \times 10^{-5}. \end{aligned}$$

Observe que el error real, 10^{-4}, excede la cota del error, $8.\overline{3} \times 10^{-5}$. Esta aparente contradicción es una consecuencia de los cálculos de dígitos finitos. Nosotros usamos la aritmética de redondeo de cuatro dígitos, y la fórmula del error de Lagrange (3.3) supone la aritmética de dígitos infinitos. Esto causó que nuestros errores reales excedieran el cálculo de error teórico.

- Recuerde: No puede esperar mayor precisión de la proporcionada por la aritmética.

El algoritmo 3.1 construye por filas las entradas en el método de Neville.

ALGORITMO **3.1**

Interpolación iterada de Neville

Para evaluar el polinomio de interpolación P en los diferentes números $n + 1$, $x_0, \ldots, x_n$ en el número x para la función f:

ENTRADA números $x, x_0, x_1, \ldots, x_n$; valores $f(x_0), f(x_1), \ldots, f(x_n)$ como la primera columna $Q_{0,0}, Q_{1,0}, \ldots, Q_{n,0}$ de Q.

SALIDA la tabla Q con $P(x) = Q_{n,n}$.

Paso 1 Para $i = 1, 2, \ldots, n$
para $j = 1, 2, \ldots, i$

haga $$Q_{i,j} = \frac{(x - x_{i-j})Q_{i,j-1} - (x - x_i)Q_{i-1,j-1}}{x_i - x_{i-j}}.$$

Paso 2 SALIDA (Q);
PARE.

La sección Conjunto de ejercicios 3.2 está disponible en línea. Encuentre la ruta de acceso en las páginas preliminares.

3.3 Diferencias divididas

La interpolación iterada se usó en la sección previa para generar sucesivamente aproximaciones polinomiales de grado superior en un punto específico. Los métodos de diferencia dividida que se presentan en esta sección se usan para generar sucesivamente los polinomios en sí mismos.

Diferencias divididas

Suponga que $P_n(x)$ es el enésimo polinomio de interpolación que concuerda con la función f en los diferentes números $x_0, x_1, \ldots, x_n$. A pesar de que este polinomio es único, existen re-

presentaciones algebraicas que son útiles en ciertas situaciones. Las diferencias divididas de f respecto a $x_0, x_1, \ldots, x_n$ se usan para expresar $P_n(x)$ en la forma

$$P_n(x) = a_0 + a_1(x - x_0) + a_2(x - x_0)(x - x_1) + \cdots + a_n(x - x_0)\cdots(x - x_{n-1}), \quad (3.5)$$

para constantes apropiadas $a_0, a_1, \ldots, a_n$. Para determinar la primera de estas constantes, a_0, observe que si $P_n(x)$ se escribe en la forma de la ecuación (3.5), entonces evaluando $P_n(x)$ en x_0 queda sólo el término constante a_0; es decir,

$$a_0 = P_n(x_0) = f(x_0).$$

Como en muchas áreas, Isaac Newton es prominente en el estudio de ecuaciones de diferencia. Desarrolló fórmulas de interpolación desde 1675, usando su notación Δ en tablas de diferencias. Adoptó un enfoque muy general hacia las fórmulas de diferencias, por lo que los ejemplos explícitos que produjo, incluyendo las fórmulas de Lagrange, a menudo son conocidas con otros nombres.

Similarmente, cuando $P(x)$ se evalúa en x_1, los únicos términos diferentes de cero en la evaluación de $P_n(x_1)$ son los términos constante y lineal,

$$f(x_0) + a_1(x_1 - x_0) = P_n(x_1) = f(x_1);$$

por lo que

$$a_1 = \frac{f(x_1) - f(x_0)}{x_1 - x_0}. \quad (3.6)$$

Ahora presentaremos la notación de diferencias divididas, que se relaciona con la notación Δ^2 de Aitkens que se usó en la sección 2.5. La *ceroésima diferencia dividida* de la función f respecto a x_i, denotada $f[x_i]$, es simplemente el valor de f en x_i:

$$f[x_i] = f(x_i). \quad (3.7)$$

Las diferencias divididas restantes se definen de manera recursiva; la *primera diferencia dividida* de f respecto a x_i y x_{i+1} se denota $f[x_i, x_{i+1}]$ y se define como

$$f[x_i, x_{i+1}] = \frac{f[x_{i+1}] - f[x_i]}{x_{i+1} - x_i}. \quad (3.8)$$

La *segunda diferencia dividida*, $f[x_i, x_{i+1}, x_{i+2}]$, se define como

$$f[x_i, x_{i+1}, x_{i+2}] = \frac{f[x_{i+1}, x_{i+2}] - f[x_i, x_{i+1}]}{x_{i+2} - x_i}.$$

De igual forma, después de que las $(k-1)$ -ésimas diferencias divididas,

$$f[x_i, x_{i+1}, x_{i+2}, \ldots, x_{i+k-1}] \quad \text{y} \quad f[x_{i+1}, x_{i+2}, \ldots, x_{i+k-1}, x_{i+k}],$$

se han determinado, la ***k*-ésima diferencia dividida** relativa a $x_i, x_{i+1}, x_{i+2}, \ldots, x_{i+k}$ es

$$f[x_i, x_{i+1}, \ldots, x_{i+k-1}, x_{i+k}] = \frac{f[x_{i+1}, x_{i+2}, \ldots, x_{i+k}] - f[x_i, x_{i+1}, \ldots, x_{i+k-1}]}{x_{i+k} - x_i}. \quad (3.9)$$

El proceso termina con la única *enésima diferencia dividida*,

$$f[x_0, x_1, \ldots, x_n] = \frac{f[x_1, x_2, \ldots, x_n] - f[x_0, x_1, \ldots, x_{n-1}]}{x_n - x_0}.$$

Debido a la ecuación (3.6), podemos escribir $a_1 = f[x_0, x_1]$, justo cuando a_0 se puede expresar como $a_0 = f(x_0) = f[x_0]$. Por lo tanto, el polinomio de interpolación en la ecuación (3.5) es

$$P_n(x) = f[x_0] + f[x_0, x_1](x - x_0) + a_2(x - x_0)(x - x_1)$$
$$+ \cdots + a_n(x - x_0)(x - x_1)\cdots(x - x_{n-1}).$$

Como se puede esperar a partir de la evaluación de a_0 y a_1, las constantes requeridas son

$$a_k = f[x_0, x_1, x_2, \ldots, x_k],$$

para cada $k = 0, 1, \ldots, n$. Por lo que $P_n(x)$, se puede reescribir en una forma llamada diferencias divididas de Newton:

$$P_n(x) = f[x_0] + \sum_{k=1}^{n} f[x_0, x_1, \ldots, x_k](x - x_0) \cdots (x - x_{k-1}). \qquad (3.10)$$

El valor de $f[x_0, x_1, \ldots, x_k]$ es independiente del orden de los números $x_0, x_1, \ldots, x_k$, como se muestra en el ejercicio 23.

La generación de las diferencias divididas se describe en la tabla 3.9. A partir de estos datos, también se pueden determinar dos cuartas y una quinta diferencia.

Tabla 3.9

x	$f(x)$	Primeras diferencias divididas	Segundas diferencias divididas	Terceras diferencias divididas
x_0	$f[x_0]$			
		$f[x_0, x_1] = \dfrac{f[x_1] - f[x_0]}{x_1 - x_0}$		
x_1	$f[x_1]$		$f[x_0, x_1, x_2] = \dfrac{f[x_1, x_2] - f[x_0, x_1]}{x_2 - x_0}$	
		$f[x_1, x_2] = \dfrac{f[x_2] - f[x_1]}{x_2 - x_1}$		$f[x_0, x_1, x_2, x_3] = \dfrac{f[x_1, x_2, x_3] - f[x_0, x_1, x_2]}{x_3 - x_0}$
x_2	$f[x_2]$		$f[x_1, x_2, x_3] = \dfrac{f[x_2, x_3] - f[x_1, x_2]}{x_3 - x_1}$	
		$f[x_2, x_3] = \dfrac{f[x_3] - f[x_2]}{x_3 - x_2}$		$f[x_1, x_2, x_3, x_4] = \dfrac{f[x_2, x_3, x_4] - f[x_1, x_2, x_3]}{x_4 - x_1}$
x_3	$f[x_3]$		$f[x_2, x_3, x_4] = \dfrac{f[x_3, x_4] - f[x_2, x_3]}{x_4 - x_2}$	
		$f[x_3, x_4] = \dfrac{f[x_4] - f[x_3]}{x_4 - x_3}$		$f[x_2, x_3, x_4, x_5] = \dfrac{f[x_3, x_4, x_5] - f[x_2, x_3, x_4]}{x_5 - x_2}$
x_4	$f[x_4]$		$f[x_3, x_4, x_5] = \dfrac{f[x_4, x_5] - f[x_3, x_4]}{x_5 - x_3}$	
		$f[x_4, x_5] = \dfrac{f[x_5] - f[x_4]}{x_5 - x_4}$		
x_5	$f[x_5]$			

ALGORITMO
3.2

Fórmula de las diferencias divididas de Newton

Para obtener los coeficientes de las diferencias divididas del polinomio de interpolación P en los $(n + 1)$ números distintos $x_0, x_1, \ldots, x_n$ para la función f:

ENTRADA los números $x_0, x_1, \ldots, x_n$; valores$f(x_0), f(x_1), \ldots, f(x_n)$ conforme $F_{0,0}, F_{1,0}, \ldots, F_{n,0}$.

SALIDA los números $F_{0,0}, F_{1,1}, \ldots, F_{n,n}$ donde

$$P_n(x) = F_{0,0} + \sum_{i=1}^{n} F_{i,i} \prod_{j=0}^{i-1} (x - x_j). \;\; (F_{i,i} \textit{ is } f[x_0, x_1, \ldots, x_i].)$$

Paso 1 Para $i = 1, 2, \ldots, n$
  Para $j = 1, 2, \ldots, i$
    haga $F_{i,j} = \dfrac{F_{i,j-1} - F_{i-1,j-1}}{x_i - x_{i-j}}$. $\quad (F_{i,j} = f[x_{i-j}, \ldots, x_i].)$

Paso 2 SALIDA $(F_{0,0}, F_{1,1}, \ldots, F_{n,n})$;
PARE. ■

La forma de la salida en el algoritmo 3.2 se puede modificar para producir todas las diferencias divididas, como se muestra en el ejemplo 1.

Ejemplo 1 Complete la tabla de diferencias divididas para los datos utilizados en el ejemplo 1 de la sección 3.2 y reproducidos en la tabla 3.10, y construya el polinomio de interpolación que usa todos estos datos.

Tabla 3.10

x	$f(x)$
1.0	0.7651977
1.3	0.6200860
1.6	0.4554022
1.9	0.2818186
2.2	0.1103623

Solución La primera diferencia dividida relacionada con x_0 y x_1 es

$$f[x_0, x_1] = \frac{f[x_1] - f[x_0]}{x_1 - x_0} = \frac{0.6200860 - 0.7651977}{1.3 - 1.0} = -0.4837057.$$

Las restantes primeras diferencias divididas se calculan de la misma forma y se muestran en la cuarta columna en la tabla 3.11

Tabla 3.11

i	x_i	$f[x_i]$	$f[x_{i-1}, x_i]$	$f[x_{i-2}, x_{i-1}, x_i]$	$f[x_{i-3}, \ldots, x_i]$	$f[x_{i-4}, \ldots, x_i]$
0	1.0	0.7651977				
			−0.4837057			
1	1.3	0.6200860		−0.1087339		
			−0.5489460		0.0658784	
2	1.6	0.4554022		−0.0494433		0.0018251
			−0.5786120		0.0680685	
3	1.9	0.2818186		0.0118183		
			−0.5715210			
4	2.2	0.1103623				

La segunda diferencia divida relacionada con x_0, x_1 y x_2 es

$$f[x_0, x_1, x_2] = \frac{f[x_1, x_2] - f[x_0, x_1]}{x_2 - x_0} = \frac{-0.5489460 - (-0.4837057)}{1.6 - 1.0} = -0.1087339.$$

Las restantes segundas diferencias divididas se muestran en la quinta columna de la tabla 3.11. La tercera diferencia dividida relacionada con x_0, x_1, x_2 y x_3 y la cuarta diferencia dividida relacionada con todos los puntos de datos son, respectivamente,

$$f[x_0, x_1, x_2, x_3] = \frac{f[x_1, x_2, x_3] - f[x_0, x_1, x_2]}{x_3 - x_0} = \frac{-0.0494433 - (-0.1087339)}{1.9 - 1.0}$$
$$= 0.0658784,$$

y

$$f[x_0, x_1, x_2, x_3, x_4] = \frac{f[x_1, x_2, x_3, x_4] - f[x_0, x_1, x_2, x_3]}{x_4 - x_0} = \frac{0.0680685 - 0.0658784}{2.2 - 1.0}$$
$$= 0.0018251.$$

Todas las entradas se dan en la tabla 3.11.

Los coeficientes de la forma de diferencias divididas hacia adelante de Newton del polinomio interpolante se encuentran a lo largo de la diagonal en la tabla. Este polinomio es

$$\begin{aligned} P_4(x) = {} & 0.7651977 - 0.4837057(x - 1.0) - 0.1087339(x - 1.0)(x - 1.3) \\ & + 0.0658784(x - 1.0)(x - 1.3)(x - 1.6) \\ & + 0.0018251(x - 1.0)(x - 1.3)(x - 1.6)(x - 1.9). \end{aligned}$$

Observe que el valor $P_4(1.5) = 0.5118200$ concuerda con el resultado en la tabla 3.6 para el ejemplo 2 de la sección 3.2, ya que los polinomios son los mismos. ■

El teorema de valor medio 1.8 aplicado a la ecuación (3.8) cuando $i = 0$,

$$f[x_0, x_1] = \frac{f(x_1) - f(x_0)}{x_1 - x_0},$$

implica que cuando existe f', $f[x_0, x_1] = f'(\xi)$ para algún número ξ entre x_0 y x_1. El siguiente teorema generaliza este resultado.

Teorema 3.6 Suponga que $f \in C^n[a, b]$ y $x_0, x_1, \ldots, x_n$ son números distintos en $[a, b]$. Entonces existe un número ξ en (a, b) con

$$f[x_0, x_1, \ldots, x_n] = \frac{f^{(n)}(\xi)}{n!}.$$

Demostración Sea

$$g(x) = f(x) - P_n(x).$$

Puesto que $f(x_i) = P_n(x_i)$ para cada $i = 0, 1, \ldots, n$, la función g tiene $n + 1$ ceros distintos en $[a, b]$. El teorema generalizado de Rolle 1.10 implica que existe un número ξ en (a, b) con $g^{(n)}(\xi) = 0$, por lo que

$$0 = f^{(n)}(\xi) - P_n^{(n)}(\xi).$$

Puesto que $P_n(x)$ es un polinomio de grado n cuyo coeficiente principal es $f[x_0, x_1, \ldots, x_n]$,

$$P_n^{(n)}(x) = n! f[x_0, x_1, \ldots, x_n],$$

para todos los valores de x. En consecuencia,

$$f[x_0, x_1, \ldots, x_n] = \frac{f^{(n)}(\xi)}{n!}.$$ ■

La fórmula de las diferencias divididas de Newton se puede expresar en forma simplificada cuando los nodos se ordenan de manera consecutiva con igual espaciado. En este caso, introducimos la notación $h = x_{i+1} - x_i$, para cada $i = 0, 1, \ldots, n-1$ y sea $x = x_0 + sh$. Entonces la diferencia $x - x_i$ es $x - x_i = (s - i)h$. Por lo que la ecuación (3.10) se convierte en

$$\begin{aligned} P_n(x) = P_n(x_0 + sh) &= f[x_0] + shf[x_0, x_1] + s(s-1)h^2 f[x_0, x_1, x_2] \\ &\quad + \cdots + s(s-1)\cdots(s-n+1)h^n f[x_0, x_1, \ldots, x_n] \\ &= f[x_0] + \sum_{k=1}^{n} s(s-1)\cdots(s-k+1)h^k f[x_0, x_1, \ldots, x_k]. \end{aligned}$$

Usando la notación de coeficiente binomial,

$$\binom{s}{k} = \frac{s(s-1)\cdots(s-k+1)}{k!},$$

podemos expresar $P_n(x)$ de manera compacta como

$$P_n(x) = P_n(x_0 + sh) = f[x_0] + \sum_{k=1}^{n} \binom{s}{k} k! h^k f[x_0, x_i, \ldots, x_k]. \tag{3.11}$$

Diferencias hacia adelante

La **fórmula de diferencias hacia adelante de Newton** se construye al usar la notación de diferencias hacia adelante Δ que se presentó en el método Δ^2 de Aitken. Con esta notación,

$$f[x_0, x_1] = \frac{f(x_1) - f(x_0)}{x_1 - x_0} = \frac{1}{h}(f(x_1) - f(x_0)) = \frac{1}{h}\Delta f(x_0)$$

$$f[x_0, x_1, x_2] = \frac{1}{2h}\left[\frac{\Delta f(x_1) - \Delta f(x_0)}{h}\right] = \frac{1}{2h^2}\Delta^2 f(x_0),$$

y, en general,

$$f[x_0, x_1, \ldots, x_k] = \frac{1}{k!h^k}\Delta^k f(x_0).$$

Puesto que $f[x_0] = f(x_0)$, la ecuación (3.11) tiene la siguiente forma.

Fórmula de diferencias hacia adelante de Newton

$$P_n(x) = f(x_0) + \sum_{k=1}^{n} \binom{s}{k} \Delta^k f(x_0) \tag{3.12}$$

Diferencias hacia atrás

Si los nodos de interpolación se reordenan desde el último hasta el primero como $x_n, x_{n-1}, \ldots, x_0$ podemos escribir la fórmula de interpolación como

$$P_n(x) = f[x_n] + f[x_n, x_{n-1}](x - x_n) + f[x_n, x_{n-1}, x_{n-2}](x - x_n)(x - x_{n-1}) + \cdots + f[x_n, \ldots, x_0](x - x_n)(x - x_{n-1}) \cdots (x - x_1).$$

Si, además, los nodos tienen el mismo espaciado con $x = x_n + sh$ y $x = x_i + (s + n - i)h$, entonces

$$\begin{aligned} P_n(x) &= P_n(x_n + sh) \\ &= f[x_n] + shf[x_n, x_{n-1}] + s(s+1)h^2 f[x_n, x_{n-1}, x_{n-2}] + \cdots \\ &\quad + s(s+1)\cdots(s+n-1)h^n f[x_n, \ldots, x_0]. \end{aligned}$$

Esto se usa para deducir una fórmula comúnmente aplicada conocida como **fórmula de diferencias hacia atrás de Newton**. Para analizar esta fórmula, necesitamos la siguiente definición.

Definición 3.7 Dada la sucesión $\{p_n\}_{n=0}^{\infty}$, defina la diferencia hacia atrás ∇p_n (lea *nabla* p_n) con

$$\nabla p_n = p_n - p_{n-1}, \quad \text{para } n \geq 1.$$

Las potencias superiores se definen de manera recursiva con

$$\nabla^k p_n = \nabla(\nabla^{k-1} p_n), \quad \text{para } k \geq 2.$$ ■

La definición 3.7 implica que

$$f[x_n, x_{n-1}] = \frac{1}{h}\nabla f(x_n), \quad f[x_n, x_{n-1}, x_{n-2}] = \frac{1}{2h^2}\nabla^2 f(x_n),$$

y, en general,

$$f[x_n, x_{n-1}, \ldots, x_{n-k}] = \frac{1}{k!h^k}\nabla^k f(x_n).$$

Por consiguiente,

$$P_n(x) = f[x_n] + s\nabla f(x_n) + \frac{s(s+1)}{2}\nabla^2 f(x_n) + \cdots + \frac{s(s+1)\cdots(s+n-1)}{n!}\nabla^n f(x_n).$$

Si extendemos la notación del coeficiente binomial para incluir todos los valores reales de s al tomar

$$\binom{-s}{k} = \frac{-s(-s-1)\cdots(-s-k+1)}{k!} = (-1)^k \frac{s(s+1)\cdots(s+k-1)}{k!},$$

entonces

$$P_n(x) = f[x_n] + (-1)^1\binom{-s}{1}\nabla f(x_n) + (-1)^2\binom{-s}{2}\nabla^2 f(x_n) + \cdots + (-1)^n\binom{-s}{n}\nabla^n f(x_n).$$

Esto nos da el siguiente resultado.

Fórmula de diferencias hacia adelante de Newton

$$P_n(x) = f[x_n] + \sum_{k=1}^{n}(-1)^k\binom{-s}{k}\nabla^k f(x_n) \qquad (3.13)$$

Ilustración La tabla 3.12 de diferencias divididas corresponde a los datos en el ejemplo 1.

Tabla 3.12

		Primeras diferencias divididas	Segundas diferencias divididas	Terceras diferencias divididas	Cuartas diferencias divididas
1.0	0.7651977				
		−0.4837057			
1.3	0.6200860		−0.1087339		
		−0.5489460		0.0658784	
1.6	0.4554022		−0.0494433		0.0018251
		−0.5786120		0.0680685	
1.9	0.2818186		0.0118183		
		−0.5715210			
2.2	0.1103623				

Solamente un polinomio de interpolación de grado, a lo sumo, cuatro, usa estos cinco puntos de datos, pero nosotros organizaremos los nodos para obtener mejores aproximaciones de interpolación de grados uno, dos y tres. Esto nos dará un sentido de precisión para la aproximación de cuarto grado para el valor dado de x.

Si se requiere una aproximación para $f(1.1)$, la opción razonable para los nodos sería $x_0 = 1.0$, $x_1 = 1.3$, $x_2 = 1.6$, $x_3 = 1.9$ y $x_4 = 2.2$, puesto que esta opción usa lo antes posible los puntos de datos más cercanos a $x = 1.1$ y también usa la cuarta diferencia dividida. Esto implica que $h = 0.3$ y $s = \frac{1}{3}$, por lo que la fórmula de diferencias dividas hacia adelante de Newton se utiliza con las diferencias divididas que tienen un subrayado *sólido* (___) en la tabla 3.12:

$$P_4(1.1) = P_4(1.0 + \frac{1}{3}(0.3))$$

$$= 0.7651977 + \frac{1}{3}(0.3)(-0.4837057) + \frac{1}{3}\left(-\frac{2}{3}\right)(0.3)^2(-0.1087339)$$

$$+ \frac{1}{3}\left(-\frac{2}{3}\right)\left(-\frac{5}{3}\right)(0.3)^3(0.0658784)$$

$$+\frac{1}{3}\left(-\frac{2}{3}\right)\left(-\frac{5}{3}\right)\left(-\frac{8}{3}\right)(0.3)^4(0.0018251)$$

$$= 0.7196460.$$ ■

Para aproximar un valor cuando x está cerca del final de los valores tabulados, digamos, $x = 2.0$, de nuevo nos gustaría usar lo antes posible los puntos de datos con $s = -\frac{2}{3}$ y las diferencias divididas en la tabla 3.12 que tienen un subrayado ondulado ($\sim\sim\sim$). Observe que la cuarta diferencia dividida se usa en ambas fórmulas:

$$\begin{aligned}
P_4(2.0) &= P_4\left(2.2 - \frac{2}{3}(0.3)\right) \\
&= 0.1103623 - \frac{2}{3}(0.3)(-0.5715210) - \frac{2}{3}\left(\frac{1}{3}\right)(0.3)^2(0.0118183) \\
&\quad - \frac{2}{3}\left(\frac{1}{3}\right)\left(\frac{4}{3}\right)(0.3)^3(0.0680685) - \frac{2}{3}\left(\frac{1}{3}\right)\left(\frac{4}{3}\right)\left(\frac{7}{3}\right)(0.3)^4(0.0018251) \\
&= 0.2238754.
\end{aligned}$$

Diferencias centradas

Las fórmulas de diferencias hacia adelante y hacia atrás de Newton no son adecuadas para aproximar $f(x)$ cuando x se encuentra cerca del centro de la tabla porque ninguna permitirá que la diferencia de orden superior tenga x_0 cerca de x. Existen varias fórmulas de diferencias divididas para este caso, cada una de las cuales incluye situaciones en las que se pueden usar para una máxima ventaja. Estos métodos reciben el nombre de **fórmulas de diferencias centradas**. Nosotros sólo consideraremos una fórmula de diferencias centradas, el método de Stirling.

Para las fórmulas de diferencias centradas seleccionamos x_0 cerca del punto que se va a aproximar y etiquetamos los nodos directamente bajo x_0 como $x_1, x_2, \ldots$ y aquellos directamente arriba como $x_{-1}, x_{-2}, \ldots$ Con esta convención la **fórmula de Stirling** está dada por

$$\begin{aligned}
P_n(x) = P_{2m+1}(x) &= f[x_0] + \frac{sh}{2}(f[x_{-1}, x_0] + f[x_0, x_1]) + s^2h^2 f[x_{-1}, x_0, x_1] \qquad (3.14) \\
&+ \frac{s(s^2-1)h^3}{2} f[x_{-2}, x_{-1}, x_0, x_1] + f[x_{-1}, x_0, x_1, x_2]) \\
&+ \cdots + s^2(s^2-1)(s^2-4)\cdots(s^2-(m-1)^2)h^{2m} f[x_{-m}, \ldots, x_m] \\
&+ \frac{s(s^2-1)\cdots(s^2-m^2)h^{2m+1}}{2}(f[x_{-m-1}, \ldots, x_m] + f[x_{-m}, \ldots, x_{m+1}]),
\end{aligned}$$

James Stirling (1692–1770) publicó ésta y muchas otras fórmulas en *Methodus Differentialis* en 1720. En este trabajo se incluyen las técnicas para acelerar la convergencia de diferentes series.

si $n = 2m + 1$ es impar. Si $n = 2m$ es par, usamos la misma fórmula pero borramos la última línea. Las entradas utilizadas para esta fórmula están subrayadas en la tabla 3.13.

Tabla 3.13

x	$f(x)$	Primeras diferencias divididas	Segundas diferencias divididas	Terceras diferencias divididas	Cuartas diferencias divididas
x_{-2}	$f[x_{-2}]$				
		$f[x_{-2}, x_{-1}]$			
x_{-1}	$f[x_{-1}]$		$f[x_{-2}, x_{-1}, x_0]$		
		$\underline{f[x_{-1}, x_0]}$		$\underline{f[x_{-2}, x_{-1}, x_0, x_1]}$	
x_0	$\underline{f[x_0]}$		$\underline{f[x_{-1}, x_0, x_1]}$		$\underline{f[x_{-2}, x_{-1}, x_0, x_1, x_2]}$
		$\underline{f[x_0, x_1]}$		$\underline{f[x_{-1}, x_0, x_1, x_2]}$	
x_1	$f[x_1]$		$f[x_0, x_1, x_2]$		
		$f[x_1, x_2]$			
x_2	$f[x_2]$				

Ejemplo 2 Considere la tabla de datos dada en los ejemplos anteriores. Use la fórmula de Stirling para aproximar $f(1.5)$ con $x_0 = 1.6$.

Solución Para aplicar la fórmula de Stirling, usamos las entradas *subrayadas* en la tabla de diferencias 3.14.

Tabla 3.14

x	$f(x)$	Primeras diferencias divididas	Segundas diferencias divididas	Terceras diferencias divididas	Cuartas diferencias divididas
1.0	0.7651977				
		−0.4837057			
1.3	0.6200860		−0.1087339		
		<u>−0.5489460</u>		<u>0.0658784</u>	
1.6	<u>0.4554022</u>		<u>−0.0494433</u>		<u>0.0018251</u>
		<u>−0.5786120</u>		<u>0.0680685</u>	
1.9	0.2818186		0.0118183		
		−0.5715210			
2.2	0.1103623				

La fórmula con, $h = 0.3, x_0 = 1.6$ y $s = -\frac{1}{3}$, se convierte en

$$f(1.5) \approx P_4\left(1.6 + \left(-\frac{1}{3}\right)(0.3)\right)$$

$$= 0.4554022 + \left(-\frac{1}{3}\right)\left(\frac{0.3}{2}\right)((-0.5489460) + (-0.5786120))$$

$$+ \left(-\frac{1}{3}\right)^2 (0.3)^2(-0.0494433)$$

$$+ \frac{1}{2}\left(-\frac{1}{3}\right)\left(\left(-\frac{1}{3}\right)^2 - 1\right)(0.3)^3(0.0658784 + 0.0680685)$$

$$+ \left(-\frac{1}{3}\right)^2\left(\left(-\frac{1}{3}\right)^2 - 1\right)(0.3)^4(0.0018251) = 0.5118200.$$

■

Muchos textos sobre análisis numérico, escritos antes del uso generalizado de las computadoras, incluyen amplios tratamientos de los métodos de diferencias divididas. Si se necesita un tratamiento más exhaustivo de este tema, el libro de Hildebrand [Hild] es una referencia especialmente buena.

La sección Conjunto de ejercicios 3.3 está disponible en línea. Encuentre la ruta de acceso en las páginas preliminares.



3.4 Interpolación de Hermite

Cuando la palabra latina *osculum*, literalmente “boca pequeña” o “beso”, se aplica a una curva, indica que sólo la toca y tiene la misma forma. La interpolación de Hermite tiene esta propiedad osculante. Corresponde a una curva dada y su derivada obliga a que la curva de interpolación “bese” a la curva dada

Los *polinomios osculantes* generalizan tanto los polinomios de Taylor como los polinomios de Lagrange. Suponga que tenemos $n + 1$ números distintos $x_0, x_1, \ldots, x_n$ en $[a, b]$ y enteros no negativos $m_0, m_1, \ldots, m_n$, y $m = \text{máx}\{m_0, m_1, \ldots, m_n\}$. El polinomio osculante que aproxima una función $f \in C^m[a, b]$ en x_i, para cada $i = 0, \ldots, n$, es el polinomio de grado mínimo que tiene los mismos valores que la función f y todas sus derivadas de orden menor que o igual que m_i en cada x_i. El grado de este polinomio osculante es el máximo

$$M = \sum_{i=0}^{n} m_i + n$$

Ya que el número de condiciones que se satisfacen es $\sum_{i=0}^{n} m_i + (n+1)$ y un polinomio de grado M tiene $M+1$ coeficientes que se pueden usar para satisfacer estas condiciones.

Definición 3.8 Sean $x_0, x_1, \ldots, x_n$ $n+1$ números distintos en $[a, b]$ y para cada $i = 0, 1, \ldots, n$, sea m_i un entero no negativo. Suponga que $f \in C^m[a, b]$, donde $m = \max_{0 \le i \le n} m_i$.

El **polinomio osculante** que se aproxima a f es el polinomio $P(x)$ de menor grado, tal que

$$\frac{d^k P(x_i)}{dx^k} = \frac{d^k f(x_i)}{dx^k}, \quad \text{para cada } i = 0, 1, \ldots, n \quad \text{y} \quad k = 0, 1, \ldots, m_i. \quad ■$$

Charles Hermite (1822–1901) realizó importantes descubrimientos matemáticos a lo largo de su vida en áreas como el análisis complejo y la teoría numérica, especialmente en relación con la teoría de las ecuaciones. Es, tal vez, mejor conocido por probar, en 1873, que *e* es trascendental; es decir, no es la solución para cualquier ecuación algebraica que tenga coeficientes enteros. Esto condujo, en 1882, a la prueba de Lindemann que establece que π también es trascendental, lo cual demostró que es imposible utilizar las herramientas de la geometría estándar de Euclides para construir un cuadrado que tenga la misma área que un círculo unitario.

Observe que cuando $n = 0$, el polinomio osculante que se aproxima a f es el m_0-ésimo polinomio de Taylor para f en x_0. Cuando $m_i = 0$ para cada i, el polinomio osculante es el enésimo polinomio de Lagrange que interpola f en $x_0, x_1, \ldots, x_n$.

Polinomios de Hermite

Cuando $m_i = 1$, para cada $i = 0, 1, \ldots, n$, nos da los **polinomios de Hermite**. Para una función f determinada, estos polinomios concuerdan con f en $x_0, x_1, \ldots, x_n$. Además, puesto que sus primeras derivadas concuerdan con las de f, tienen la misma "forma" que la función en $(x_i, f(x_i))$, en el sentido en el que las *rectas tangentes* al polinomio y la función concuerdan. Nosotros limitaremos nuestro estudio de los polinomios osculantes a esta situación y primero consideraremos un teorema que describe de manera precisa la forma de los polinomios de Hermite.

Teorema 3.9 Si $f \in C^1[a, b]$ y $x_0, \ldots, x_n \in [a, b]$ son distintos, el único polinomio de menor grado que concuerda con f y f' en $x_0, \ldots, x_n$ es el polinomio de Hermite de grado a lo sumo $2n+1$ dado por

$$H_{2n+1}(x) = \sum_{j=0}^{n} f(x_j) H_{n,j}(x) + \sum_{j=0}^{n} f'(x_j) \hat{H}_{n,j}(x),$$

Donde cada $L_{n,j}(x)$ denota el j-ésimo coeficiente del polinomio de Lagrange de grado n, y

$$H_{n,j}(x) = [1 - 2(x - x_j) L'_{n,j}(x_j)] L^2_{n,j}(x) \quad \text{y} \quad \hat{H}_{n,j}(x) = (x - x_j) L^2_{n,j}(x).$$

Además, si $f \in C^{2n+2}[a, b]$, entonces

$$f(x) = H_{2n+1}(x) + \frac{(x - x_0)^2 \ldots (x - x_n)^2}{(2n+2)!} f^{(2n+2)}(\xi(x)),$$

para algunos (en general desconocidos) $\xi(x)$ en el intervalo (a, b). ■

En 1878, Hermite dio una descripción de un polinomio osculante general en una carta para Carl W. Borchardt, a quien regularmente enviaba sus resultados nuevos. Su demostración es una aplicación interesante del uso de técnicas complejas de integración para resolver un problema de valor real.

Demostración Primero, recuerde que

$$L_{n,j}(x_i) = \begin{cases} 0, & \text{si } i \ne j, \\ 1, & \text{si } i = j. \end{cases}$$

Por lo tanto, cuando $i \ne j$,

$$H_{n,j}(x_i) = 0 \quad \text{y} \quad \hat{H}_{n,j}(x_i) = 0,$$

Mientras que, para cada i,

$$H_{n,i}(x_i) = [1 - 2(x_i - x_i) L'_{n,i}(x_i)] \cdot 1 = 1 \quad \text{y} \quad \hat{H}_{n,i}(x_i) = (x_i - x_i) \cdot 1^2 = 0.$$

Por consiguiente,

$$H_{2n+1}(x_i) = \sum_{\substack{j=0 \\ j\neq i}}^{n} f(x_j)\cdot 0 + f(x_i)\cdot 1 + \sum_{j=0}^{n} f'(x_j)\cdot 0 = f(x_i),$$

de modo que H_{2n+1} concuerda con f en $x_0, x_1, \ldots, x_n$.

Para mostrar la concordancia de H'_{2n+1} con f' en los nodos, primero observe que $L_{n,j}(x)$ es un factor de $H'_{n,j}(x)$, por lo que $H'_{n,j}(x_i) = 0$ cuando $i \neq j$ Además, cuando $i = j$, tenemos $L_{n,i}(x_i) = 1$, por lo que

$$\begin{aligned} H'_{n,i}(x_i) &= -2L'_{n,i}(x_i)\cdot L^2_{n,i}(x_i) + [1 - 2(x_i - x_i)L'_{n,i}(x_i)]2L_{n,i}(x_i)L'_{n,i}(x_i) \\ &= -2L'_{n,i}(x_i) + 2L'_{n,i}(x_i) = 0. \end{aligned}$$

Por lo tanto, $H'_{n,j}(x_i) = 0$ para todas las i y j.

Finalmente,

$$\begin{aligned} \hat{H}'_{n,j}(x_i) &= L^2_{n,j}(x_i) + (x_i - x_j)2L_{n,j}(x_i)L'_{n,j}(x_i) \\ &= L_{n,j}(x_i)[L_{n,j}(x_i) + 2(x_i - x_j)L'_{n,j}(x_i)], \end{aligned}$$

por lo que $\hat{H}'_{n,j}(x_i) = 0$ si $i \neq j$ y $\hat{H}'_{n,i}(x_i) = 1$. Al combinar estos hechos, tenemos

$$H'_{2n+1}(x_i) = \sum_{j=0}^{n} f(x_j)\cdot 0 + \sum_{\substack{j=0 \\ j\neq i}}^{n} f'(x_j)\cdot 0 + f'(x_i)\cdot 1 = f'(x_i).$$

Por lo tanto, H_{2n+1} concuerda con f y H'_{2n+1} con f' en $x_0, x_1, \ldots, x_n$.

La unicidad de este polinomio y la deducción de la fórmula del error se consideran en el ejercicio 11. ■

Ejemplo 1 Use el polinomio de Hermite que concuerda con los datos listados en la tabla 3.5 para encontrar una aproximación de $f(1.5)$.

Tabla 3.15

k	x_k	$f(x_k)$	$f'(x_k)$
0	1.3	0.6200860	−0.5220232
1	1.6	0.4554022	−0.5698959
2	1.9	0.2818186	−0.5811571

Solución Primero calculamos los polinomios de Lagrange y sus derivadas. Esto nos da

$$L_{2,0}(x) = \frac{(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)} = \frac{50}{9}x^2 - \frac{175}{9}x + \frac{152}{9}, \qquad L'_{2,0}(x) = \frac{100}{9}x - \frac{175}{9};$$

$$L_{2,1}(x) = \frac{(x-x_0)(x-x_2)}{(x_1-x_0)(x_1-x_2)} = \frac{-100}{9}x^2 + \frac{320}{9}x - \frac{247}{9}, \quad L'_{2,1}(x) = \frac{-200}{9}x + \frac{320}{9};$$

y

$$L_{2,2} = \frac{(x-x_0)(x-x_1)}{(x_2-x_0)(x_2-x_1)} = \frac{50}{9}x^2 - \frac{145}{9}x + \frac{104}{9}, \qquad L'_{2,2}(x) = \frac{100}{9}x - \frac{145}{9}.$$

Los polinomios $H_{2,j}(x)$ y $\hat{H}_{2,j}(x)$ son entonces

$$H_{2,0}(x) = [1 - 2(x - 1.3)(-5)]\left(\frac{50}{9}x^2 - \frac{175}{9}x + \frac{152}{9}\right)^2$$

$$= (10x - 12)\left(\frac{50}{9}x^2 - \frac{175}{9}x + \frac{152}{9}\right)^2,$$

$$H_{2,1}(x) = 1 \cdot \left(\frac{-100}{9}x^2 + \frac{320}{9}x - \frac{247}{9}\right)^2,$$

$$H_{2,2}(x) = 10(2 - x)\left(\frac{50}{9}x^2 - \frac{145}{9}x + \frac{104}{9}\right)^2,$$

$$\hat{H}_{2,0}(x) = (x - 1.3)\left(\frac{50}{9}x^2 - \frac{175}{9}x + \frac{152}{9}\right)^2,$$

$$\hat{H}_{2,1}(x) = (x - 1.6)\left(\frac{-100}{9}x^2 + \frac{320}{9}x - \frac{247}{9}\right)^2,$$

y

$$\hat{H}_{2,2}(x) = (x - 1.9)\left(\frac{50}{9}x^2 - \frac{145}{9}x + \frac{104}{9}\right)^2.$$

Finalmente,

$$H_5(x) = 0.6200860H_{2,0}(x) + 0.4554022H_{2,1}(x) + 0.2818186H_{2,2}(x)$$
$$- 0.5220232\hat{H}_{2,0}(x) - 0.5698959\hat{H}_{2,1}(x) - 0.5811571\hat{H}_{2,2}(x)$$

y

$$H_5(1.5) = 0.6200860\left(\frac{4}{27}\right) + 0.4554022\left(\frac{64}{81}\right) + 0.2818186\left(\frac{5}{81}\right)$$
$$- 0.5220232\left(\frac{4}{405}\right) - 0.5698959\left(\frac{-32}{405}\right) - 0.5811571\left(\frac{-2}{405}\right) = 0.5118277,$$

un resultado que es apropiado para los lugares enumerados. ■

A pesar de que el teorema 3.9 proporciona una descripción completa de los polinomios de Hermite, a partir del ejemplo 1 es claro que la necesidad de determinar y evaluar los polinomios de Lagrange y sus derivadas hace que el procedimientos sea tedioso para los valores pequeños de n.

Polinomios de Hermite usando diferencias divididas

Existe un método alterno para generar aproximaciones de Hermite que tiene sus bases en la fórmula de diferencias divididas de interpolación de Newton (3.10) en $x_0, x_1, \ldots, x_n$; esto es,

$$P_n(x) = f[x_0] + \sum_{k=1}^{n} f[x_0, x_1, \ldots, x_k](x - x_0)\cdots(x - x_{k-1}).$$

El método alterno utiliza la conexión entre la enésima diferencia dividida y la enésima derivada de f, como se describe en el teorema 3.6 en la sección 3.3.

Suponga que los diferentes números $x_0, x_1, \ldots, x_n$ están dados junto con los valores de f y f' en estos números. Defina una nueva sucesión $z_0, z_1, \ldots, z_{2n+1}$ mediante

$$z_{2i} = z_{2i+1} = x_i, \quad \text{para cada } i = 0, 1, \ldots, n,$$

y construya la tabla de diferencias divididas en la forma de la tabla 3.9 que usa z_0, z_1, …, z_{2n+1}.

Puesto que $z_{2i} = z_{2i+1} = x_i$ para cada i, no podemos definir $f[z_{2i}, z_{2i+1}]$ con la fórmula de diferencias divididas. Sin embargo, si suponemos, con base en el teorema 3.6, que la sustitución razonable en estas situaciones es $f[z_{2i}, z_{2i+1}] = f'(z_{2i}) = f'(x_i)$, podemos usar las entradas

$$f'(x_0), f'(x_1), \dots, f'(x_n)$$

en lugar de las primeras diferencias divididas no definidas

$$f[z_0, z_1], f[z_2, z_3], \dots, f[z_{2n}, z_{2n+1}].$$

Las diferencias divididas restantes se producen de la manera común y las diferencias divididas adecuadas se usan en la fórmula de diferencias divididas de interpolación de Newton. La tabla 3.16 muestra las entradas que se utilizan para las primeras tres columnas de diferencias divididas al determinar el polinomio de Hermite $H_5(x)$ para x_0, x_1 y x_2. Las entradas restantes se generan de la manera que se muestra en la tabla 3.9. El polinomio de Hermite está dado por

$$H_{2n+1}(x) = f[z_0] + \sum_{k=1}^{2n+1} f[z_0, \dots, z_k](x - z_0)(x - z_1) \cdots (x - z_{k-1}).$$

Una prueba de este hecho puede encontrarse en [Pow], p. 56

Tabla 3.16

z	$f(z)$	Primeras diferencias divididas	Segundas diferencias divididas
$z_0 = x_0$	$f[z_0] = f(x_0)$		
		$f[z_0, z_1] = f'(x_0)$	
$z_1 = x_0$	$f[z_1] = f(x_0)$		$f[z_0, z_1, z_2] = \dfrac{f[z_1, z_2] - f[z_0, z_1]}{z_2 - z_0}$
		$f[z_1, z_2] = \dfrac{f[z_2] - f[z_1]}{z_2 - z_1}$	
$z_2 = x_1$	$f[z_2] = f(x_1)$		$f[z_1, z_2, z_3] = \dfrac{f[z_2, z_3] - f[z_1, z_2]}{z_3 - z_1}$
		$f[z_2, z_3] = f'(x_1)$	
$z_3 = x_1$	$f[z_3] = f(x_1)$		$f[z_2, z_3, z_4] = \dfrac{f[z_3, z_4] - f[z_2, z_3]}{z_4 - z_2}$
		$f[z_3, z_4] = \dfrac{f[z_4] - f[z_3]}{z_4 - z_3}$	
$z_4 = x_2$	$f[z_4] = f(x_2)$		$f[z_3, z_4, z_5] = \dfrac{f[z_4, z_5] - f[z_3, z_4]}{z_5 - z_3}$
		$f[z_4, z_5] = f'(x_2)$	
$z_5 = x_2$	$f[z_5] = f(x_2)$		

Ejemplo 2 Use los datos que se proporcionan en el ejemplo 1 y el método de diferencias divididas para determinar la aproximación polinomial de Hermite en $x = 1.5$.

Solución Las entradas subrayadas en las primeras tres columnas de la tabla 3.17 son los datos que se proporcionaron en el ejemplo 1. Las entradas restantes en esta tabla se generan con la fórmula de diferencias divididas estándar (3.9).

Por ejemplo, para la segunda entrada en la tercera columna usamos la segunda entrada 1.3 en la segunda columna y la primera entrada 1.6 en esa columna para obtener

$$\frac{0.4554022 - 0.6200860}{1.6 - 1.3} = -0.5489460.$$

Para la primera entrada en la cuarta columna, usamos la primera entrada 1.3 en la tercera columna y la primera entrada 1.6 en esa columna para obtener

$$\frac{-0.5489460-(-0.5220232)}{1.6-1.3}=-0.0897427.$$

El valor del polinomio de Hermite en 1.5 es

$$\begin{aligned}
H_5(1.5) = {} & f[1.3]+f'(1.3)(1.5-1.3)+f[1.3,1.3,1.6](1.5-1.3)^2\\
& + f[1.3,1.3,1.6,1.6](1.5-1.3)^2(1.5-1.6)\\
& + f[1.3,1.3,1.6,1.6,1.9](1.5-1.3)^2(1.5-1.6)^2\\
& + f[1.3,1.3,1.6,1.6,1.9,1.9](1.5-1.3)^2(1.5-1.6)^2(1.5-1.9)\\
= {} & 0.6200860+(-0.5220232)(0.2)+(-0.0897427)(0.2)^2\\
& + 0.0663657(0.2)^2(-0.1)+0.0026663(0.2)^2(-0.1)^2\\
& + (-0.0027738)(0.2)^2(-0.1)^2(-0.4)\\
= {} & 0.5118277.
\end{aligned}$$

■

Tabla 3.17

1.3	0.6200860					
		−0.5220232				
1.3	0.6200860		−0.0897427			
		−0.5489460		0.0663657		
1.6	0.4554022		−0.0698330		0.0026663	
		−0.5698959		0.0679655		−0.0027738
1.6	0.4554022		−0.0290537		0.0010020	
		−0.5786120		0.0685667		
1.9	0.2818186		−0.0084837			
		−0.5811571				
1.9	0.2818186					

La técnica que se usa en el algoritmo 3.3 se puede ampliar para su uso en la determinación de otros polinomios osculantes. Es posible encontrar un análisis conciso de los procedimientos en [Pow], pp. 53–57.

ALGORITMO 3.3

Interpolación de Hermite

Para obtener los coeficientes del polinomio de interpolación de Hermite $H(x)$ en los $(n+1)$ números distintos $x_0, \ldots, x_n$ para la función f:

ENTRADA números $x_0, x_1, \ldots, x_n$; valores $f(x_0), \ldots, f(x_n)$ y $f'(x_0), \ldots, f'(x_n)$.

SALIDA los números $Q_{0,0}, Q_{1,1}, \ldots, Q_{2n+1,2n+1}$ donde

$$\begin{aligned}
H(x) = {} & Q_{0,0}+Q_{1,1}(x-x_0)+Q_{2,2}(x-x_0)^2+Q_{3,3}(x-x_0)^2(x-x_1)\\
& + Q_{4,4}(x-x_0)^2(x-x_1)^2+\cdots\\
& + Q_{2n+1,2n+1}(x-x_0)^2(x-x_1)^2\cdots(x-x_{n-1})^2(x-x_n).
\end{aligned}$$

Paso 1 Para $i=0,1,\ldots,n$ haga los pasos 2 y 3.

Paso 2 Haga $z_{2i}=x_i$;
$z_{2i+1}=x_i$;
$Q_{2i,0}=f(x_i)$;
$Q_{2i+1,0}=f(x_i)$;
$Q_{2i+1,1}=f'(x_i)$.

Paso 3 Si $i \neq 0$ entonces haga

$$Q_{2i,1} = \frac{Q_{2i,0} - Q_{2i-1,0}}{z_{2i} - z_{2i-1}}.$$

Paso 4 Para $i = 2, 3, \ldots, 2n+1$

para $j = 2, 3, \ldots, i$ haga $Q_{i,j} = \dfrac{Q_{i,j-1} - Q_{i-1,j-1}}{z_i - z_{i-j}}$.

Paso 5 SALIDA $(Q_{0,0}, Q_{1,1}, \ldots, Q_{2n+1,2n+1})$;
PARE. ■

3.5 Interpolación de spline cúbico[1]

Las secciones previas se preocuparon por la aproximación de las funciones arbitrarias en intervalos cerrados usando un polinomio individual. Sin embargo, los polinomios de orden superior pueden oscilar erráticamente; es decir, una fluctuación menor sobre una pequeña parte del intervalo puede inducir fluctuaciones grandes sobre todo el rango. Observaremos un buen ejemplo de esto en la figura 3.14 al final de esta sección.

Un enfoque alternativo es dividir el intervalo de aproximación en un conjunto de subintervalos y construir (en general) un polinomio de aproximación diferente en cada subintervalo. Esto se llama **aproximación polinomial por tramos**.

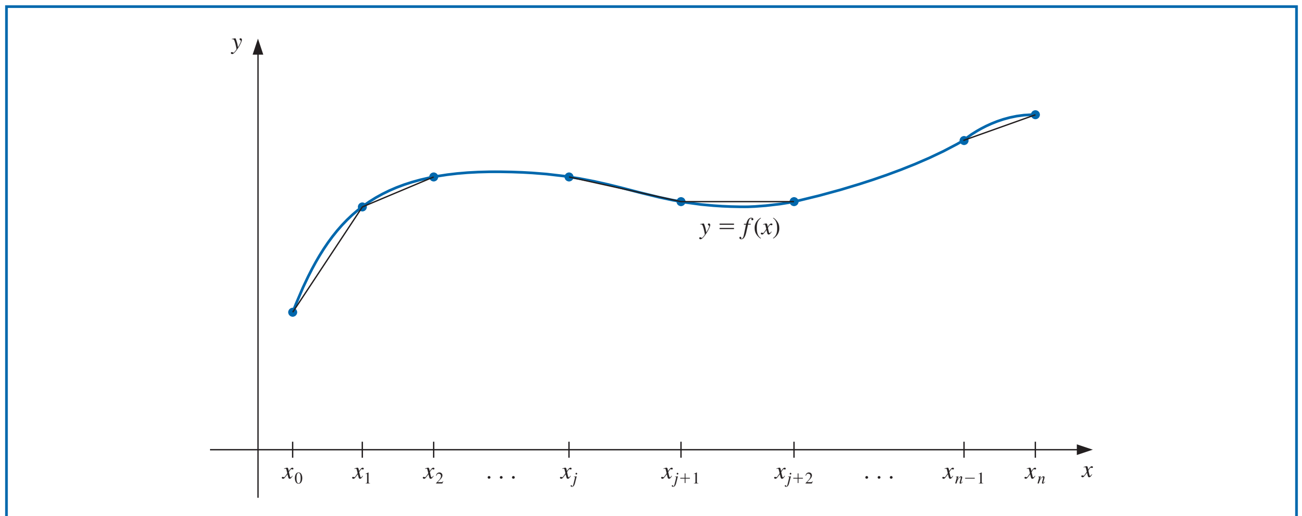
Aproximación de polinomio por tramos

La aproximación polinomial por tramos más simple es la interpolación **lineal por tramos**, la cual consiste en unir un conjunto de puntos de datos

$$\{(x_0, f(x_0)), (x_1, f(x_1)), \ldots, (x_n, f(x_n))\}$$

mediante una serie de líneas rectas, como se muestra en la figura 3.7.

Figura 3.7



Una desventaja de la aproximación mediante funciones lineales es que probablemente no existe diferenciabilidad en los extremos de los subintervalos, lo que, en un contexto

[1]Las pruebas de los teoremas en esta sección dependen de los resultados en el capítulo 6.

geométrico, significa que la función de interpolación no es "suave". A menudo es claro, a partir de las condiciones físicas, que se requiere suavidad, por lo que la función de aproximación debe ser continuamente diferenciable.

Un procedimiento alterno es usar un polinomio por tramos de tipo Hermite. Por ejemplo, si se conocen los valores de f y de f' en cada uno de los puntos $x_0 < x_1 < \ldots < x_n$, se puede usar un polinomio cúbico de Hermite en cada uno de los subintervalos $[x_0, x_1], [x_1, x_2], \ldots, [x_{n-1}, x_n]$ para obtener una función que tiene una derivada continua en un intervalo $[x_0, x_n]$.

Determinar el polinomio cúbico de Hermite adecuado en un intervalo dado es simplemente cuestión de calcular $H_3(x)$ para ese intervalo. Los polinomios de interpolación de Lagrange necesarios para determinar H_3, son de primer grado, por lo que esto se puede lograr sin mayor dificultad. Sin embargo, para usar los polinomios por tramos de Hermite para la interpolación general, necesitamos conocer la derivada de la función que se va a aproximar y esto con frecuencia no está disponible.

Isaac Jacob Schoenberg (1903–1990) desarrolló su trabajo sobre splines durante la Segunda Guerra Mundial mientras estaba de licencia en la Universidad de Pennsylvania para trabajar en el Army's Ballistic Research Laboratory (Laboratorio de Investigación de Balística del Ejército) en Aberdeen, Maryland. Su trabajo original implicaba procedimientos numéricos para resolver ecuaciones diferenciales. La aplicación mucho más amplia de los splines en las áreas de ajuste de datos y diseño de geometría asistida por computador se volvieron evidentes con la disponibilidad generalizada de las computadoras en la década de 1960.

El resto de esta sección considera que la aproximación usa polinomios por tramos que no requieren información específica sobre la derivada excepto, tal vez, en los extremos del intervalo en el que la función se aproxima.

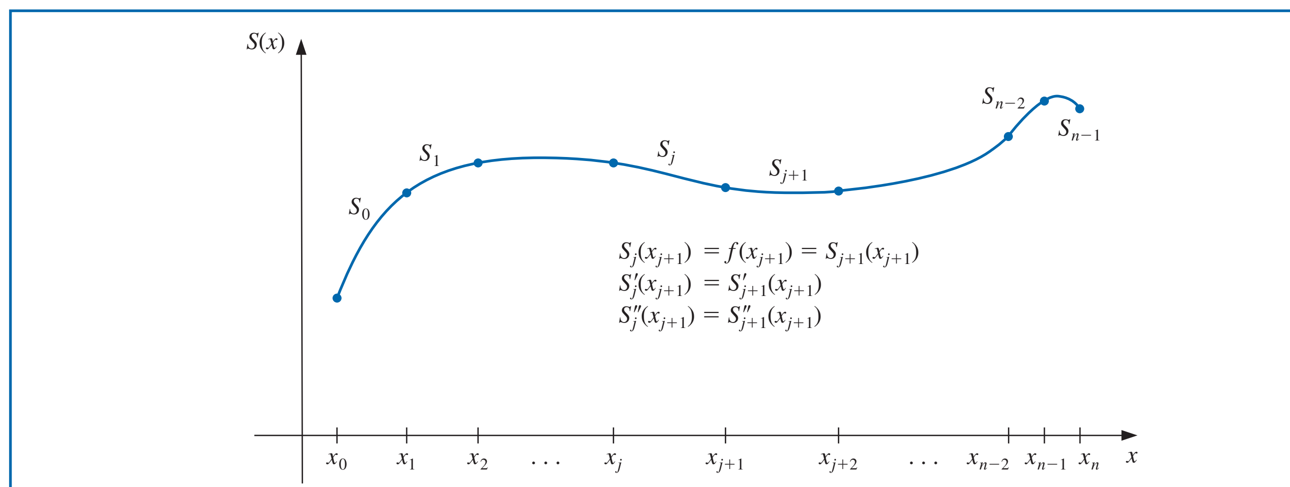
El tipo más simple de función polinomial por tramos diferenciable en un intervalo completo $[x_0, x_1]$ es la función obtenida al ajustar un polinomio cuadrático entre cada par sucesivo de nodos. Esto se hace al construir una cuadrática en $[x_0, x_1]$ que concuerda con la función en x_0 y x_1, otra cuadrática en $[x_1, x_2]$ que concuerda con la función en x_1 y x_2, y así sucesivamente. Un polinomio cuadrático general tiene tres constantes arbitrarias: el término constante, el coeficiente de x y el coeficiente de x^2, y sólo se requieren dos condiciones para ajustar los datos en los extremos de cada subintervalo. Por lo tanto, existe flexibilidad que permite seleccionar las cuadráticas de tal forma que el interpolante tenga una derivada continua en $[x_0, x_n]$. La dificultad surge porque generalmente necesitamos especificar condiciones sobre la derivada del interpolante en los extremos x_0 y x_n. No hay un número suficiente de constantes para garantizar que las condiciones se satisfagan (consulte el ejercicio 34).

La raíz de la palabra "spline" es la misma que la de "splint". Originalmente, era una tira pequeña de madera que se puede utilizar para unir dos tablas. Más adelante, la palabra se usó para dibujar curvas suaves y continuas al forzar que la tira pasara a través de puntos específicos y siguiera a lo largo de la curva.

Splines cúbicos

La aproximación polinomial por tramos más común usa polinomios cúbicos entre cada par sucesivo de nodos y recibe el nombre de **interpolación de spline cúbico**. Un polinomio cúbico general implica cuatro constantes, por lo que existe suficiente flexibilidad en el procedimiento de spline cúbico para garantizar que el interpolante no sólo es continuamente diferenciable en el intervalo, sino también tiene una segunda derivada continua. Sin embargo, la construcción del spline cúbico no supone que las derivadas del interpolante concuerdan con las de la función en su aproximación, incluso en los nodos (consulte la figura 3.8.)

Figura 3.8



Definición 3.10 Dada una función f definida en $[a, b]$ y un conjunto de nodos $a = x_0 < x_1 < \cdots < x_n = b$, un **interpolante de spline cúbico** S para f es una función que satisface las siguientes condiciones:

a) $S(x)$ es un polinomio cúbico, que se denota $S_j(x)$, en el subintervalo $[x_j, x_{j+1}]$ para cada $j = 0, 1, \ldots, n-1$;

b) $S_j(x_j) = f(x_j)$ y $S_j(x_{j+1}) = f(x_{j+1})$ para cada $j = 0, 1, \ldots, n-1$;

c) $S_{j+1}(x_{j+1}) = S_j(x_{j+1})$ para cada $j = 0, 1, \ldots, n-2$; (*implícito en* **b)**.)

d) $S'_{j+1}(x_{j+1}) = S'_j(x_{j+1})$ para cada $j = 0, 1, \ldots, n-2$;

e) $S''_{j+1}(x_{j+1}) = S''_j(x_{j+1})$ para cada $j = 0, 1, \ldots, n-2$;

f) Uno de los siguientes conjuntos de condiciones de frontera se satisface:

i) $S''(x_0) = S''(x_n) = 0$ **(frontera natural)** *(o libre)*;

ii) $S'(x_0) = f'(x_0)$ y $S'(x_n) = f'(x_n)$ **(frontera condicionada)**.

Un spline natural no tiene condiciones impuestas para la dirección en sus extremos, por lo que la curva toma la forma de una línea recta después de pasar por los puntos de interpolación más cercanos a sus extremos. El nombre deriva del hecho de que ésta es la forma natural que asume una tira flexible, si es forzada a pasar por los puntos de interpolación específicos sin restricciones adicionales (consulte la figura 3.9.)



Figura 3.9

Aunque los splines cúbicos se definen con otras condiciones de frontera, las condiciones dadas en la parte f) son suficientes para nuestros propósitos. Cuando se presentan condiciones de frontera libres, el spline recibe el nombre de **spline natural** y su gráfica se aproxima a la forma que una varilla flexible y larga asumiría si fuera forzada a pasar por los puntos de datos $\{(x_0, f(x_0)), (x_1, f(x_1)), \ldots, (x_n, f(x_n))\}$.

En general, las condiciones de frontera condicionada conducen a aproximaciones más precisas porque incluyen más información sobre la función. Sin embargo, para mantener este tipo de condición de frontera, es necesario tener, ya sea los valores de la derivada en los extremos o una aproximación precisa para esos valores.

Ejemplo 1 Construya un spline cúbico natural que pase por los puntos (1, 2), (2, 3) y (3, 5).

Solución Este spline consiste en dos cúbicos. El primero para el intervalo [1, 2], que se denota

$$S_0(x) = a_0 + b_0(x-1) + c_0(x-1)^2 + d_0(x-1)^3,$$

y el otro para [2, 3], que se denota

$$S_1(x) = a_1 + b_1(x-2) + c_1(x-2)^2 + d_1(x-2)^3.$$

Existen ocho constantes que se van a determinar y esto requiere ocho condiciones. Cuatro condiciones a partir del hecho de que los splines deben concordar con los datos en los nodos. Por lo tanto,

$$2 = f(1) = a_0, \quad 3 = f(2) = a_0 + b_0 + c_0 + d_0, \quad 3 = f(2) = a_1, \quad \text{y}$$
$$5 = f(3) = a_1 + b_1 + c_1 + d_1.$$

Dos más provienen del hecho de que $S'_0(2) = S'_1(2)$ y $S''_0(2) = S''_1(2)$. Estos son

$$S'_0(2) = S'_1(2): \quad b_0 + 2c_0 + 3d_0 = b_1 \qquad \text{y} \qquad S''_0(2) = S''_1(2): \quad 2c_0 + 6d_0 = 2c_1.$$

Los dos finales provienen de las condiciones de frontera natural:

$$S''_0(1) = 0: \quad 2c_0 = 0 \qquad \text{y} \qquad S''_1(3) = 0: \quad 2c_1 + 6d_1 = 0.$$

Al resolver este sistema de ecuaciones obtenemos el spline

$$S(x) = \begin{cases} 2 + \frac{3}{4}(x-1) + \frac{1}{4}(x-1)^3, & \text{para } x \in [1, 2] \\ 3 + \frac{3}{2}(x-2) + \frac{3}{4}(x-2)^2 - \frac{1}{4}(x-2)^3, & \text{para } x \in [2, 3]. \end{cases}$$

■

Construcción de un spline cúbico

Como lo demuestra el ejemplo previo, un spline definido en un intervalo que se ha dividido en n subintervalos requerirá determinar $4n$ constantes. Para construir el spline cúbico que se interpola para una función dada f, las condiciones en la definición se aplican a los polinomios cúbicos

$$S_j(x) = a_j + b_j(x - x_j) + c_j(x - x_j)^2 + d_j(x - x_j)^3,$$

para cada $j = 0, 1, \ldots, n - 1$. Puesto que $S_j(x_j) = a_j = f(x_j)$, la condición **c**) se puede aplicar para obtener

$$a_{j+1} = S_{j+1}(x_{j+1}) = S_j(x_{j+1}) = a_j + b_j(x_{j+1} - x_j) + c_j(x_{j+1} - x_j)^2 + d_j(x_{j+1} - x_j)^3,$$

para cada $j = 0, 1, \ldots, n - 2$.

Sujetar un spline indica que los extremos de la tira flexible están fijos de tal forma que cada uno de sus extremos es forzado a tomar una dirección específica. Esto es importante, por ejemplo, cuando los extremos de dos funciones spline deben concordar. Esto se realiza de manera matemática al especificar los valores de la derivada de la curva en los extremos del spline.

Como los términos $x_{j+1} - x_j$ son usados repetidamente en este desarrollo, es conveniente introducir una notación más simple

$$h_j = x_{j+1} - x_j,$$

para cada $j = 0, 1, \ldots, n - 1$. Si también definimos $a_n = f(x_n)$, entonces la ecuación

$$a_{j+1} = a_j + b_j h_j + c_j h_j^2 + d_j h_j^3 \tag{3.15}$$

se mantiene para cada $j = 0, 1, \ldots, n - 1$.

De manera similar, defina $b_n = S'(x_n)$ y observe que

$$S_j'(x) = b_j + 2c_j(x - x_j) + 3d_j(x - x_j)^2$$

implica que $S_j'(x_j) = b_j$, para *cada* $j = 0, 1, \ldots, n - 1$. Al aplicar la condición en la parte d) obtenemos

$$b_{j+1} = b_j + 2c_j h_j + 3d_j h_j^2, \tag{3.16}$$

para cada $j = 0, 1, \ldots, n - 1$.

Otra relación entre los coeficientes de S_j se obtiene al definir $c_n = S''(x_n)/2$ y aplicar la condición en la parte e). Entonces, para cada $j = 0, 1, \ldots, n - 1$,

$$c_{j+1} = c_j + 3d_j h_j. \tag{3.17}$$

Resolviendo para d_j en la ecuación (3.17) y sustituyendo este valor en las ecuaciones (3.15) y (3.16) obtenemos, para cada $j = 0, 1, \ldots, n - 1$, las nuevas ecuaciones

$$a_{j+1} = a_j + b_j h_j + \frac{h_j^2}{3}(2c_j + c_{j+1}) \tag{3.18}$$

y

$$b_{j+1} = b_j + h_j(c_j + c_{j+1}). \tag{3.19}$$

La relación final que involucra los coeficientes se obtiene al resolver la ecuación adecuada en la forma de la ecuación (3.18), primero para b_j,

$$b_j = \frac{1}{h_j}(a_{j+1} - a_j) - \frac{h_j}{3}(2c_j + c_{j+1}), \tag{3.20}$$

y entonces, con una reducción del índice, para b_{j-1}. Esto nos da

$$b_{j-1} = \frac{1}{h_{j-1}}(a_j - a_{j-1}) - \frac{h_{j-1}}{3}(2c_{j-1} + c_j).$$

Al sustituir estos valores en la ecuación obtenida de la ecuación (3.19), con el índice reducido en uno, obtenemos el sistema lineal de ecuaciones

$$h_{j-1}c_{j-1} + 2(h_{j-1} + h_j)c_j + h_j c_{j+1} = \frac{3}{h_j}(a_{j+1} - a_j) - \frac{3}{h_{j-1}}(a_j - a_{j-1}), \quad (3.21)$$

Para cada $j = 1, 2, \ldots, n-1$. Este sistema sólo tiene los $\{c_j\}_{j=0}^{n}$ como incógnitas. Los valores de $\{h_j\}_{j=0}^{n-1}$ y $\{a_j\}_{j=0}^{n}$ están dados, respectivamente, por el espaciado de los nodos $\{x_j\}_{j=0}^{n}$ y los valores de f en los nodos. Por lo que, una vez que se determinan los valores de $\{c_j\}_{j=0}^{n}$, es sencillo encontrar el resto de las constantes $\{b_j\}_{j=0}^{n-1}$ a partir de la ecuación (3.20) y $\{d_j\}_{j=0}^{n-1}$ a partir de la ecuación (3.17). Entonces podemos construir los polinomios cúbicos $\{S_j(x)\}_{j=0}^{n-1}$.

La pregunta más importante que surge en relación con esta construcción es si los valores de $\{c_j\}_{j=0}^{n}$ se pueden encontrar usando el sistema de ecuaciones dado en la ecuación (3.21) y, en este caso, si estos valores son únicos. Los siguientes teoremas indican que éste es el caso cuando se impone cualquiera de las condiciones de frontera dadas en la parte f) de la definición. Las demostraciones de estos teoremas requieren material a partir del álgebra lineal, la cual se analiza en el capítulo 6.

Splines naturales

Teorema 3.11 Si f se define en $a = x_0 < x_1 < \cdots < x_n = b$, entonces f tiene un spline natural único que interpola S en los nodos $x_0, x_1, \ldots, x_n$; es decir, un spline interpolante que satisface las condiciones de frontera natural $S''(a) = 0$ y $S''(b) = 0$.

Demostración Las condiciones de frontera en este caso implican que $c_n = S''(x_n)/2 = 0$ y que

$$0 = S''(x_0) = 2c_0 + 6d_0(x_0 - x_0),$$

por lo que $c_0 = 0$. Las dos ecuaciones $c_0 = 0$ y $c_n = 0$ junto con las ecuaciones en (3.21) producen un sistema lineal descrito por la ecuación matriz-vector $A\mathbf{x} = \mathbf{b}$, donde A es la matriz $(n+1) \times (n+1)$

$$A = \begin{bmatrix} 1 & 0 & 0 & \cdots & \cdots & \cdots & 0 \\ h_0 & 2(h_0 + h_1) & h_1 & \ddots & & & \vdots \\ 0 & h_1 & 2(h_1 + h_2) & h_2 & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & h_{n-2} & 2(h_{n-2} + h_{n-1}) & & h_{n-1} \\ 0 & \cdots & \cdots & 0 & 0 & & 1 \end{bmatrix},$$

y **b** y **x** son los vectores

$$\mathbf{b} = \begin{bmatrix} 0 \\ \frac{3}{h_1}(a_2 - a_1) - \frac{3}{h_0}(a_1 - a_0) \\ \vdots \\ \frac{3}{h_{n-1}}(a_n - a_{n-1}) - \frac{3}{h_{n-2}}(a_{n-1} - a_{n-2}) \\ 0 \end{bmatrix} \quad \text{y} \quad \mathbf{x} = \begin{bmatrix} c_0 \\ c_1 \\ \vdots \\ c_n \end{bmatrix}.$$

La matriz A es estrictamente dominante de manera diagonal; es decir, en cada fila, la magnitud de la entrada diagonal excede la suma de las magnitudes de todas las otras entradas en la fila. Un sistema lineal con una matriz de esta forma se mostrará mediante el teorema 6.21 en la sección 6.6 para tener una única solución para $c_0, c_1, \ldots, c_n$. ■

La solución para el problema del spline cúbico con las condiciones de frontera $S''(x_0) = S''(x_n) = 0$ se puede obtener al aplicar el algoritmo 3.4.

ALGORITMO **3.4**

Spline cúbico natural

Para construir el spline cúbico interpolante S para la función f, definido en los números $x_0 < x_1 < \cdots < x_n$, que satisfacen $S''(x_0) = S''(x_n) = 0$:

ENTRADA $n; x_0, x_1, \ldots, x_n; a_0 = f(x_0), a_1 = f(x_1), \ldots, a_n = f(x_n)$.

SALIDA a_j, b_j, c_j, d_j para $j = 0, 1, \ldots, n-1$.
(*Nota:* $S(x) = S_j(x) = a_j + b_j(x - x_j) + c_j(x - x_j)^2 + d_j(x - x_j)^3$ *para* $x_j \leq x \leq x_{j+1}$.)

Paso 1 Para $i = 0, 1, \ldots, n-1$ haga $h_i = x_{i+1} - x_i$.

Paso 2 Para $i = 1, 2, \ldots, n-1$ haga

$$\alpha_i = \frac{3}{h_i}(a_{i+1} - a_i) - \frac{3}{h_{i-1}}(a_i - a_{i-1}).$$

Paso 3 Determine $l_0 = 1$; (*Los pasos 3, 4 y 5 y parte del paso 6 resuelven un sistema lineal tridiagonal con un método descrito en el algoritmo 6.7.*)
$\mu_0 = 0$;
$z_0 = 0$.

Paso 4 Para $i = 1, 2, \ldots, n-1$
haga $l_i = 2(x_{i+1} - x_{i-1}) - h_{i-1}\mu_{i-1}$;
$\mu_i = h_i / l_i$;
$z_i = (\alpha_i - h_{i-1}z_{i-1})/l_i$.

Paso 5 Haga $l_n = 1$;
$z_n = 0$;
$c_n = 0$.

Paso 6 Para $j = n-1, n-2, \ldots, 0$
haga $c_j = z_j - \mu_j c_{j+1}$;
$b_j = (a_{j+1} - a_j)/h_j - h_j(c_{j+1} + 2c_j)/3$;
$d_j = (c_{j+1} - c_j)/(3h_j)$.

Paso 7 SALIDA $(a_j, b_j, c_j, d_j$ para $j = 0, 1, \ldots, n-1)$;
PARE. ■

Ejemplo 2 Al inicio del capítulo 3, proporcionamos algunos polinomios de Taylor para aproximar la exponencial $f(x) = e^x$. Use los puntos $(0, 1)$, $(1, e)$, $(2, e^2)$, y $(3, e^3)$ para formar un spline natural $S(x)$ que se aproxima a $f(x) = e^x$.

Solución Tenemos $n = 3$, $h_0 = h_1 = h_2 = 1$, $a_0 = 1$, $a_1 = e$, $a_2 = e^2$, y $a_3 = e^3$. Por lo que, la matriz A y los vectores **b** y **x** determinados en el teorema 3.11 tienen las formas

$$A = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 4 & 1 & 0 \\ 0 & 1 & 4 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} 0 \\ 3(e^2 - 2e + 1) \\ 3(e^3 - 2e^2 + e) \\ 0 \end{bmatrix}, \quad \text{y} \quad \mathbf{x} = \begin{bmatrix} c_0 \\ c_1 \\ c_2 \\ c_3 \end{bmatrix}.$$

La ecuación matriz-vector $A\mathbf{x} = \mathbf{b}$ es equivalente al sistema de ecuaciones

$$c_0 = 0,$$

$$c_0 + 4c_1 + c_2 = 3(e^2 - 2e + 1),$$

$$c_1 + 4c_2 + c_3 = 3(e^3 - 2e^2 + e),$$

$$c_3 = 0.$$

Este sistema tiene la solución $c_0 = c_3 = 0$, y para cinco lugares decimales,

$$c_1 = \frac{1}{5}(-e^3 + 6e^2 - 9e + 4) \approx 0.75685, \quad \text{y} \quad c_2 = \frac{1}{5}(4e^3 - 9e^2 + 6e - 1) \approx 5.83007.$$

Al resolver para las constantes restantes obtenemos

$$b_0 = \frac{1}{h_0}(a_1 - a_0) - \frac{h_0}{3}(c_1 + 2c_0)$$

$$= (e - 1) - \frac{1}{15}(-e^3 + 6e^2 - 9e + 4) \approx 1.46600,$$

$$b_1 = \frac{1}{h_1}(a_2 - a_1) - \frac{h_1}{3}(c_2 + 2c_1)$$

$$= (e^2 - e) - \frac{1}{15}(2e^3 + 3e^2 - 12e + 7) \approx 2.22285,$$

$$b_2 = \frac{1}{h_2}(a_3 - a_2) - \frac{h_2}{3}(c_3 + 2c_2)$$

$$= (e^3 - e^2) - \frac{1}{15}(8e^3 - 18e^2 + 12e - 2) \approx 8.80977,$$

$$d_0 = \frac{1}{3h_0}(c_1 - c_0) = \frac{1}{15}(-e^3 + 6e^2 - 9e + 4) \approx 0.25228,$$

$$d_1 = \frac{1}{3h_1}(c_2 - c_1) = \frac{1}{3}(e^3 - 3e^2 + 3e - 1) \approx 1.69107,$$

y

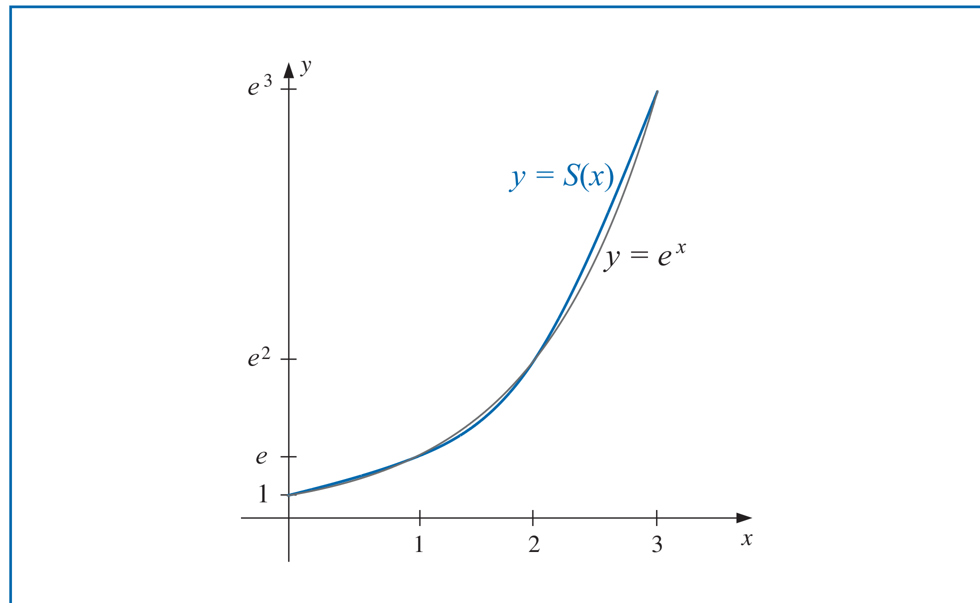
$$d_2 = \frac{1}{3h_2}(c_3 - c_1) = \frac{1}{15}(-4e^3 + 9e^2 - 6e + 1) \approx -1.94336.$$

El spline cúbico natural se describe por tramos mediante

$$S(x)=\begin{cases}1+1.46600x+0.25228x^3, & \text{para } x\in[0,1],\\ 2.71828+2.22285(x-1)+0.75685(x-1)^2+1.69107(x-1)^3, & \text{para } x\in[1,2],\\ 7.38906+8.80977(x-2)+5.83007(x-2)^2-1.94336(x-2)^3, & \text{para } x\in[2,3].\end{cases}$$

El spline y $f(x) = e^x$ se muestran en la figura 3.10. ■

Figura 3.10



Una vez que hemos determinado un spline para una aproximación de una función, podemos usarlo para aproximar otras propiedades de la función. La siguiente ilustración implica la integral del spline que encontramos en el ejemplo previo.

Ilustración Para aproximar la integral de $f(x) = e^x$ en $[0, 3]$, que tiene el valor

$$\int_0^3 e^x\,dx = e^3 - 1 \approx 20.08553692 - 1 = 19.08553692,$$

podemos integrar por tramos el spline que aproxima f en este intervalo. Esto nos da

$$\begin{aligned}\int_0^3 S(x) = &\int_0^1 1 + 1.46600x + 0.25228x^3\,dx\\ &+\int_1^2 2.71828 + 2.22285(x-1) + 0.75685(x-1)^2 + 1.69107(x-1)^3\,dx\\ &+\int_2^3 7.38906 + 8.80977(x-2) + 5.83007(x-2)^2 - 1.94336(x-2)^3\,dx.\end{aligned}$$

Integrando y calculando los valores de las potencias obtenemos

$$\int_0^3 S(x) = \left[x + 1.46600\frac{x^2}{2} + 0.25228\frac{x^4}{4}\right]_0^1$$

$$+ \left[2.71828(x-1) + 2.22285\frac{(x-1)^2}{2} + 0.75685\frac{(x-1)^3}{3} + 1.69107\frac{(x-1)^4}{4}\right]_1^2$$

$$+ \left[7.38906(x-2) + 8.80977\frac{(x-2)^2}{2} + 5.83007\frac{(x-2)^3}{3} - 1.94336\frac{(x-2)^4}{4}\right]_2^3$$

$$= (1 + 2.71828 + 7.38906) + \frac{1}{2}(1.46600 + 2.22285 + 8.80977)$$

$$+ \frac{1}{3}(0.75685 + 5.83007) + \frac{1}{4}(0.25228 + 1.69107 - 1.94336)$$

$$= 19.55229.$$

Puesto que los nodos están espaciados de manera equivalente en este ejemplo, la aproximación de la integral es simplemente

$$\int_0^3 S(x)\,dx = (a_0 + a_1 + a_2) + \frac{1}{2}(b_0 + b_1 + b_2) + \frac{1}{3}(c_0 + c_1 + c_2) + \frac{1}{4}(d_0 + d_1 + d_2). \tag{3.22}$$

■

Splines condicionados

Ejemplo 3 En el ejemplo 1 encontramos un spline natural S que pasa por los puntos (1, 2), (2, 3) y (3, 5). Construir un spline condicionado s que pase por esos puntos y que cumpla $s'(1) = 2$ y $s'(3) = 1$.

Solución Si

$$s_0(x) = a_0 + b_0(x-1) + c_0(x-1)^2 + d_0(x-1)^3$$

es el cúbico en $[1, 2]$ y el cúbico en $[2, 3]$ es

$$s_1(x) = a_1 + b_1(x-2) + c_1(x-2)^2 + d_1(x-2)^3.$$

Entonces, la mayoría de las condiciones para determinar ocho constantes son iguales a las del ejemplo 1. Es decir,

$$2 = f(1) = a_0, \quad 3 = f(2) = a_0 + b_0 + c_0 + d_0, \quad 3 = f(2) = a_1, \quad \text{y}$$
$$5 = f(3) = a_1 + b_1 + c_1 + d_1.$$

$$s_0'(2) = s_1'(2): \quad b_0 + 2c_0 + 3d_0 = b_1 \quad \text{y} \quad s_0''(2) = s_1''(2): \quad 2c_0 + 6d_0 = 2c_1$$

Sin embargo, las condiciones de frontera ahora son

$$s_0'(1) = 2: \quad b_0 = 2 \quad \text{y} \quad s_1'(3) = 1: \quad b_1 + 2c_1 + 3d_1 = 1.$$

Resolviendo este sistema de ecuaciones obtenemos el spline como

$$s(x) = \begin{cases} 2 + 2(x-1) - \frac{5}{2}(x-1)^2 + \frac{3}{2}(x-1)^3, & \text{para } x \in [1, 2] \\ 3 + \frac{3}{2}(x-2) + 2(x-2)^2 - \frac{3}{2}(x-2)^3, & \text{para } x \in [2, 3] \end{cases}.$$ ■

En el caso general de las condiciones de frontera condicionada, tenemos un resultado que es similar al teorema para las condiciones de frontera natural descritas en el teorema 3.11.

Teorema 3.12 Si f se define en $a = x_0 < x_1 < \cdots < x_n = b$ y es diferenciable en a y b, entonces f tiene un único spline interpolante S condicionado en los nodos $x_0, x_1, \ldots, x_n$; es decir, un spline interpolante que satisface las condiciones de frontera condicionada $S'(a) = f'(a)$ y $S'(b) = f'(b)$.

Demostración Puesto que $f'(a) = S'(a) = S'(x_0) = b_0$, la ecuación (3.20) con $j = 0$ implica que

$$f'(a) = \frac{1}{h_0}(a_1 - a_0) - \frac{h_0}{3}(2c_0 + c_1).$$

Por consiguiente,

$$2h_0c_0 + h_0c_1 = \frac{3}{h_0}(a_1 - a_0) - 3f'(a).$$

De igual forma,

$$f'(b) = b_n = b_{n-1} + h_{n-1}(c_{n-1} + c_n),$$

por lo que la ecuación (3.20) con $j = n - 1$ implica que

$$\begin{aligned} f'(b) &= \frac{a_n - a_{n-1}}{h_{n-1}} - \frac{h_{n-1}}{3}(2c_{n-1} + c_n) + h_{n-1}(c_{n-1} + c_n) \\ &= \frac{a_n - a_{n-1}}{h_{n-1}} + \frac{h_{n-1}}{3}(c_{n-1} + 2c_n), \end{aligned}$$

y

$$h_{n-1}c_{n-1} + 2h_{n-1}c_n = 3f'(b) - \frac{3}{h_{n-1}}(a_n - a_{n-1}).$$

Las ecuaciones (3.21) junto con las ecuaciones

$$2h_0c_0 + h_0c_1 = \frac{3}{h_0}(a_1 - a_0) - 3f'(a)$$

y

$$h_{n-1}c_{n-1} + 2h_{n-1}c_n = 3f'(b) - \frac{3}{h_{n-1}}(a_n - a_{n-1})$$

determinan el sistema lineal $A\mathbf{x} = \mathbf{b}$, donde

$$A = \begin{bmatrix} 2h_0 & h_0 & 0 & \cdots & \cdots & 0 \\ h_0 & 2(h_0+h_1) & h_1 & \ddots & & \vdots \\ 0 & h_1 & 2(h_1+h_2) & h_2 & \ddots & 0 \\ \vdots & \ddots & \ddots & h_{n-2} & 2(h_{n-2}+h_{n-1}) & h_{n-1} \\ 0 & \cdots & \cdots & 0 & h_{n-1} & 2h_{n-1} \end{bmatrix},$$

$$\mathbf{b} = \begin{bmatrix} \frac{3}{h_0}(a_1 - a_0) - 3f'(a) \\ \frac{3}{h_1}(a_2 - a_1) - \frac{3}{h_0}(a_1 - a_0) \\ \vdots \\ \frac{3}{h_{n-1}}(a_n - a_{n-1}) - \frac{3}{h_{n-2}}(a_{n-1} - a_{n-2}) \\ 3f'(b) - \frac{3}{h_{n-1}}(a_n - a_{n-1}) \end{bmatrix}, \quad \text{y} \quad \mathbf{x} = \begin{bmatrix} c_0 \\ c_1 \\ \vdots \\ c_n \end{bmatrix}.$$

Esta matriz A también es estrictamente dominante de manera diagonal, por lo que satisface las condiciones del teorema 6.21 en la sección 6.6. Por lo tanto, el sistema lineal tiene una solución única para $c_0, c_1, \ldots, c_n$. ■

La solución del problema de spline cúbico con condiciones de frontera $S'(x_0) = f'(x_0)$ y $S'(x_n) = f'(x_n)$ se puede obtener al aplicar el algoritmo 3.5.

ALGORITMO 3.5

Spline cúbico condicionado

Para construir el spline cúbico interpolante S para la función f definida en los números $x_0 < x_1 < \cdots < x_n$, que satisfacen $S'(x_0) = f'(x_0)$ y $S'(x_n) = f'(x_n)$:

ENTRADA n; $x_0, x_1, \ldots, x_n$; $a_0 = f(x_0), a_1 = f(x_1), \ldots, a_n = f(x_n)$; $FPO = f'(x_0)$; $FPN = f'(x_n)$.

SALIDA a_j, b_j, c_j, d_j para $j = 0, 1, \ldots, n-1$.
(*Nota:* $S(x) = S_j(x) = a_j + b_j(x - x_j) + c_j(x - x_j)^2 + d_j(x - x_j)^3$ *para* $x_j \le x \le x_{j+1}$.)

Paso 1 Para $i = 0, 1, \ldots, n-1$ haga $h_i = x_{i+1} - x_i$.

Paso 2 Haga $\alpha_0 = 3(a_1 - a_0)/h_0 - 3FPO$;
$\alpha_n = 3FPN - 3(a_n - a_{n-1})/h_{n-1}$.

Paso 3 Para $i = 1, 2, \ldots, n-1$

haga $\alpha_i = \dfrac{3}{h_i}(a_{i+1} - a_i) - \dfrac{3}{h_{i-1}}(a_i - a_{i-1})$.

Paso 4 Haga $l_0 = 2h_0$; (*Los pasos 4, 5 y 6 y parte del paso 7 resuelven un sistema lineal tridiagonal con un método descrito en el algoritmo 6.7.*)
$\mu_0 = 0.5$;
$z_0 = \alpha_0/l_0$.

Paso 5 Para $i = 1, 2, \ldots, n-1$
haga $l_i = 2(x_{i+1} - x_{i-1}) - h_{i-1}\mu_{i-1}$;
$\mu_i = h_i/l_i$;
$z_i = (\alpha_i - h_{i-1}z_{i-1})/l_i$.

Paso 6 Haga $l_n = h_{n-1}(2 - \mu_{n-1})$;
$z_n = (\alpha_n - h_{n-1}z_{n-1})/l_n$;
$c_n = z_n$.

Paso 7 Para $j = n-1, n-2, \ldots, 0$
haga $c_j = z_j - \mu_j c_{j+1}$;
$b_j = (a_{j+1} - a_j)/h_j - h_j(c_{j+1} + 2c_j)/3$;
$d_j = (c_{j+1} - c_j)/(3h_j)$.

Paso 8 SALIDA (a_j, b_j, c_j, d_j para $j = 0, 1, \ldots, n-1$);
PARE. ■

Ejemplo 4 En el ejemplo 2 se usó un spline natural y los puntos de datos (0, 1), (1, e), (2, e^2) y (3, e^3) para formar una nueva función de aproximación $S(x)$. Determine el spline condicionado $s(x)$ que utiliza estos datos y la información adicional, $f'(x) = e^x$, $f'(0) = 1$ y $f'(3) = e^3$.

Solución Como en el ejemplo 2, tenemos $n = 3, h_0 = h_1 = h_2 = 1, a_0 = 0, a_1 = e$, $a_2 = e^2$, y $a_3 = e^3$ Esto, junto con la información que $f'(0) = 1$ y $f'(3) = e^3$, da la matriz A y los vectores **b** y **x** con las formas

$$A = \begin{bmatrix} 2 & 1 & 0 & 0 \\ 1 & 4 & 1 & 0 \\ 0 & 1 & 4 & 1 \\ 0 & 0 & 1 & 2 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} 3(e-2) \\ 3(e^2 - 2e + 1) \\ 3(e^3 - 2e^2 + e) \\ 3e^2 \end{bmatrix}, \quad \text{y} \quad \mathbf{x} = \begin{bmatrix} c_0 \\ c_1 \\ c_2 \\ c_3 \end{bmatrix}.$$

La ecuación matriz-vector $A\mathbf{x} = \mathbf{b}$ es equivalente al sistema de ecuaciones

$$2c_0 + c_1 = 3(e-2),$$
$$c_0 + 4c_1 + c_2 = 3(e^2 - 2e + 1),$$
$$c_1 + 4c_2 + c_3 = 3(e^3 - 2e^2 + e),$$
$$c_2 + 2c_3 = 3e^2.$$

Resolviendo este sistema simultáneamente para c_0, c_1, c_2 y c_3 obtenemos, con cinco lugares decimales,

$$c_0 = \frac{1}{15}(2e^3 - 12e^2 + 42e - 59) = 0.44468,$$
$$c_1 = \frac{1}{15}(-4e^3 + 24e^2 - 39e + 28) = 1.26548,$$
$$c_2 = \frac{1}{15}(14e^3 - 39e^2 + 24e - 8) = 3.35087,$$
$$c_3 = \frac{1}{15}(-7e^3 + 42e^2 - 12e + 4) = 9.40815.$$

Al resolver para las constantes restantes de la misma manera que en el ejemplo 2 obtenemos

$$b_0 = 1.00000, \quad b_1 = 2.71016, \quad b_2 = 7.32652,$$

y

$$d_0 = 0.27360, \quad d_1 = 0.69513, \quad d_2 = 2.01909.$$

Esto nos da el spline cúbico condicionado

$$s(x) = \begin{cases} 1 + x + 0.44468x^2 + 0.27360x^3, & \text{si } 0 \leq x < 1, \\ 2.71828 + 2.71016(x-1) + 1.26548(x-1)^2 + 0.69513(x-1)^3, & \text{si } 1 \leq x < 2, \\ 7.38906 + 7.32652(x-2) + 3.35087(x-2)^2 + 2.01909(x-2)^3, & \text{si } 2 \leq x \leq 3. \end{cases}$$

La gráfica del spline condicionado y $f(x) = e^x$ son tan similares que no es posible observar diferencias. ■

También podemos aproximar la integral de f en $[0, 3]$ al integrar el spline condicionado. El valor exacto de la integral es

$$\int_0^3 e^x \, dx = e^3 - 1 \approx 20.08554 - 1 = 19.08554.$$

Puesto que los datos están igualmente espaciados, integrar por tramos el spline condicionado resulta en la misma fórmula que en (3.22); es decir,

$$\int_0^3 s(x) \, dx = (a_0 + a_1 + a_2) + \frac{1}{2}(b_0 + b_1 + b_2) + \frac{1}{3}(c_0 + c_1 + c_2) + \frac{1}{4}(d_0 + d_1 + d_2).$$

Por lo tanto, la aproximación de integral es

$$\begin{aligned} \int_0^3 s(x) \, dx &= (1 + 2.71828 + 7.38906) + \frac{1}{2}(1 + 2.71016 + 7.32652) \\ &\quad + \frac{1}{3}(0.44468 + 1.26548 + 3.35087) + \frac{1}{4}(0.27360 + 0.69513 + 2.01909) \\ &= 19.05965. \end{aligned}$$

El error absoluto en la aproximación integral usando los splines naturales y condicionados es

$$\text{Natural: } |19.08554 - 19.55229| = 0.46675$$

y

$$\text{Condicionado: } |19.08554 - 19.05965| = 0.02589.$$

Para propósitos de integración, el spline condicionado es inmensamente superior. Esto no debería ser una sorpresa porque las condiciones de frontera para el spline condicionado son exactas, mientras que para el spline natural fundamentalmente asumimos que, $f''(x) = e^x$,

$$0 = S''(x) \approx f''(0) = e^1 = 1 \quad \text{y} \quad 0 = S''(3) \approx f''(3) = e^3 \approx 20.$$

La siguiente ilustración usa un spline para aproximar una curva que no tiene una representación funcional.

Ilustración La figura 3.11 muestra un pato malvasia en vuelo. Para aproximar el perfil superior del pato hemos seleccionado puntos a lo largo de la curva por los cuales queremos que pase la curva de aproximación. La tabla 3.18 enumera las coordenadas de 21 puntos de datos relativos al sistema de coordenadas superpuesto que se muestra en la figura 3.12. Observe que se usan más puntos cuando la curva cambia rápidamente que cuando lo hace más despacio.

Figura 3.11

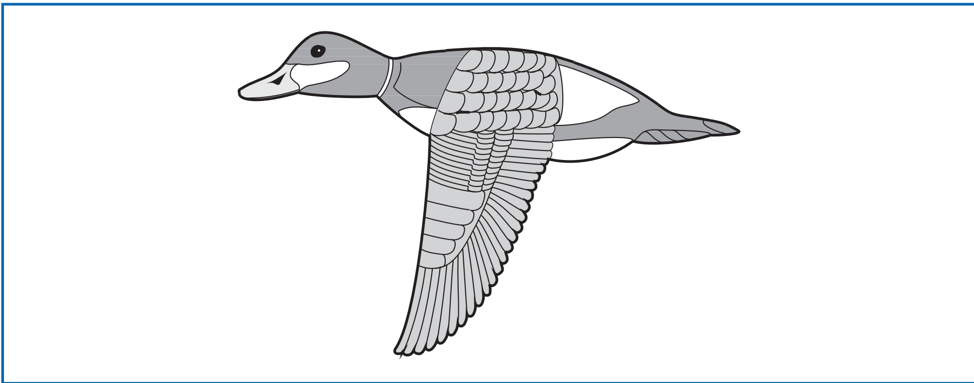
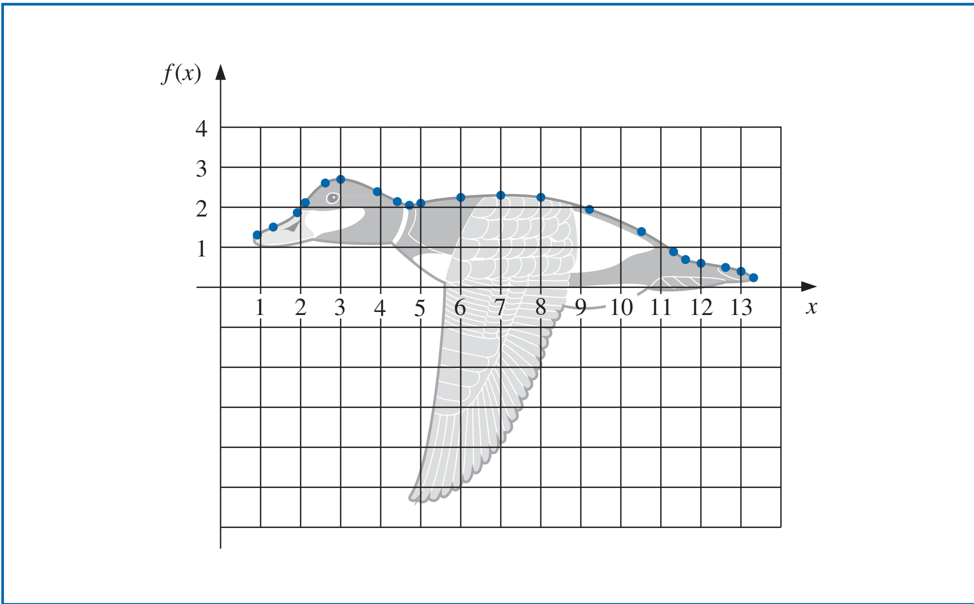


Tabla 3.18

x	0.9	1.3	1.9	2.1	2.6	3.0	3.9	4.4	4.7	5.0	6.0	7.0	8.0	9.2	10.5	11.3	11.6	12.0	12.6	13.0	13.3
$f(x)$	1.3	1.5	1.85	2.1	2.6	2.7	2.4	2.15	2.05	2.1	2.25	2.3	2.25	1.95	1.4	0.9	0.7	0.6	0.5	0.4	0.25

Figura 3.12

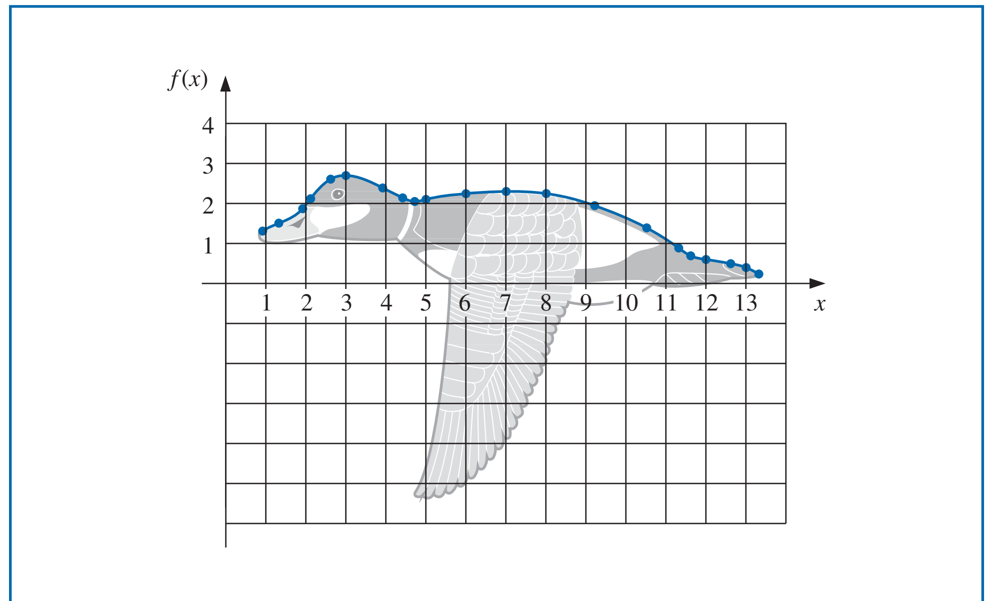


El uso del algoritmo 3.4 para generar el spline cúbico natural para estos datos produce los coeficientes que se muestran en la tabla 3.19. Esta curva spline es casi idéntica al perfil, como se muestra en la figura 3.13.

Tabla 3.19

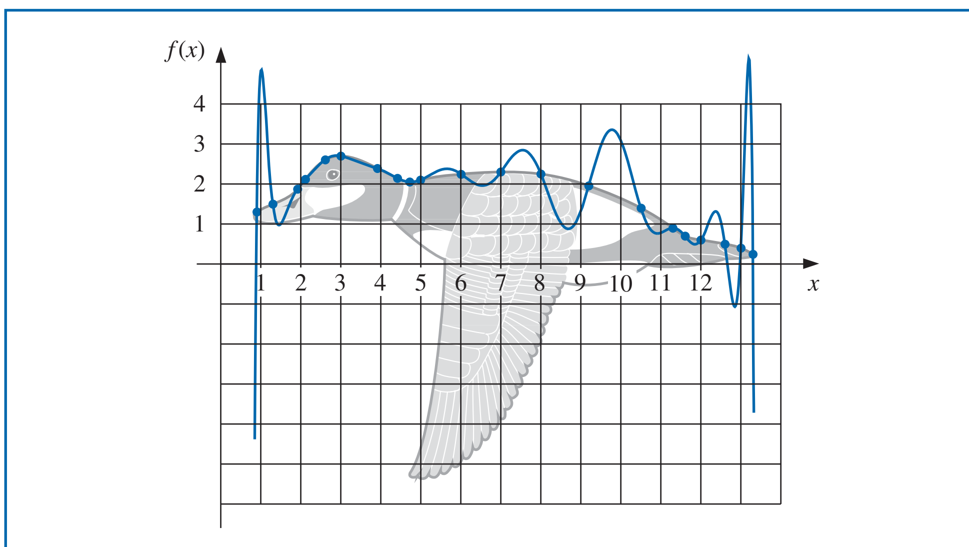
j	x_j	a_j	b_j	c_j	d_j
0	0.9	1.3	0.54	0.00	−0.25
1	1.3	1.5	0.42	−0.30	0.95
2	1.9	1.85	1.09	1.41	−2.96
3	2.1	2.1	1.29	−0.37	−0.45
4	2.6	2.6	0.59	−1.04	0.45
5	3.0	2.7	−0.02	−0.50	0.17
6	3.9	2.4	−0.50	−0.03	0.08
7	4.4	2.15	−0.48	0.08	1.31
8	4.7	2.05	−0.07	1.27	−1.58
9	5.0	2.1	0.26	−0.16	0.04
10	6.0	2.25	0.08	−0.03	0.00
11	7.0	2.3	0.01	−0.04	−0.02
12	8.0	2.25	−0.14	−0.11	0.02
13	9.2	1.95	−0.34	−0.05	−0.01
14	10.5	1.4	−0.53	−0.10	−0.02
15	11.3	0.9	−0.73	−0.15	1.21
16	11.6	0.7	−0.49	0.94	−0.84
17	12.0	0.6	−0.14	−0.06	0.04
18	12.6	0.5	−0.18	0.00	−0.45
19	13.0	0.4	−0.39	−0.54	0.60
20	13.3	0.25			

Figura 3.13



Para propósitos de comparación, la figura 3.14 proporciona una ilustración de la curva que se genera con un polinomio de interpolación de Lagrange para ajustar los datos provistos en la tabla 3.18. En este caso, el polinomio de interpolación es de grado 20 y oscila en forma desordenada. Produce una ilustración muy extraña de la espalda del pato, en vuelo o de otra forma.

Figura 3.14



Al utilizar un spline condicionado para aproximar esta curva, necesitaríamos derivar aproximaciones para los extremos. Incluso si estas aproximaciones están disponibles, podríamos esperar poca mejora debido al acuerdo cerrado del spline cúbico natural para la curva del perfil superior. ■

Construir un spline cúbico para aproximar el perfil inferior del pato malvasía sería más difícil porque la curva para esta parte no se puede expresar como una función de x, y en ciertos puntos la curva no parece ser suave. Estos problemas se pueden resolver usando splines separados para representar varias partes de la curva, pero en la siguiente sección se considera un enfoque más eficaz para aproximar las curvas de este tipo.

En general, al aproximar funciones mediante splines cúbicos son preferibles las condiciones de frontera condicionada, por lo que la derivada de la función debe conocerse o aproximarse en los extremos del intervalo. Cuando los nodos están espaciados uniformemente cerca de ambos extremos, es posible obtener las aproximaciones con cualquiera de las fórmulas adecuadas provistas en las secciones 4.1 y 4.2. Cuando los nodos no están espaciados de manera uniforme, el problema es considerablemente más difícil.

Para concluir esta sección, listamos una fórmula de la cota de error para el spline cúbico sujeto a condiciones de frontera. La prueba de este resultado se puede encontrar en [Schul], pp. 57–58.

Teorema 3.13 Sea $f \in C^4[a, b]$ con $\text{máx}_{a \le x \le b} |f^{(4)}(x)| = M$. Si S es el único spline cúbico condicionado interpolante para f respecto a los nodos $a = x_0 < x_1 < \cdots < x_n = b$, entonces, para todas las x en $[a, b]$.

$$|f(x) - S(x)| \le \frac{5M}{384} \max_{0 \le j \le n-1} (x_{j+1} - x_j)^4.$$ ■

Un resultado de la cota del error de cuarto orden también se mantiene para el caso de las condiciones de frontera natural, pero es más difícil de expresar (consulte [BD], pp. 827–835).

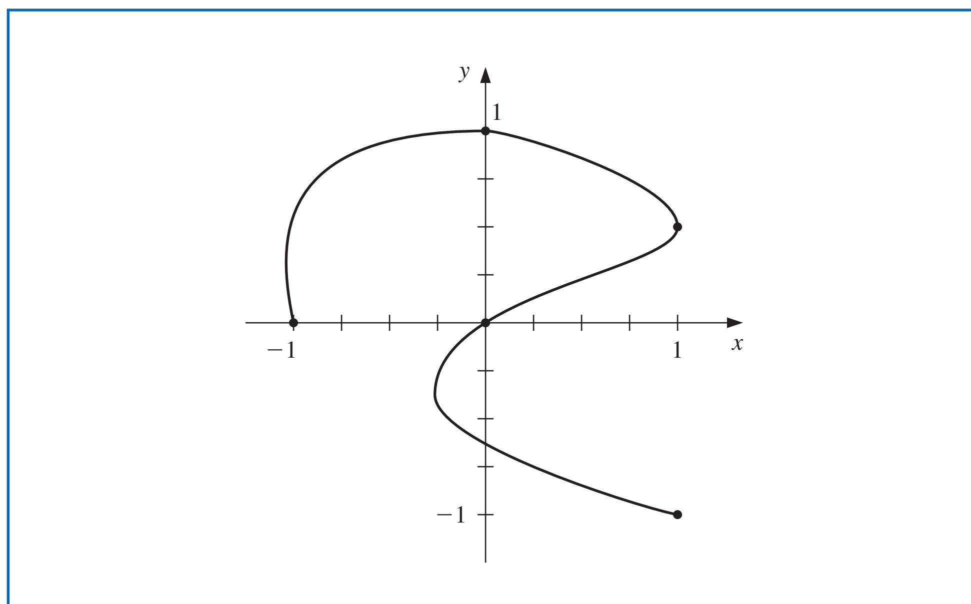
En general, las condiciones de frontera natural proporcionarán resultados menos precisos que las de frontera condicionada cerca de los extremos del intervalo $[x_0, x_n]$ a menos que la función f satisfaga $f''(x_0) = f''(x_n) = 0$. Una alternativa para la condición de frontera natural que no requiere conocimiento de la derivada de f es la condición *sin nudo* (consulte [Deb2], pp. 55–56). Esta condición requiere que $S'''(x)$ sea continua en x_1 y x_{n-1}.

La sección Conjunto de ejercicios 3.5 está disponible en línea. Encuentre la ruta de acceso en las páginas preliminares.

3.6 Curvas paramétricas

Ninguna de las técnicas desarrolladas en este capítulo puede usarse para generar curvas de la forma que se muestra en la figura 3.15, porque esta curva no se puede expresar como una función de una variable coordenada en términos de la otra. En esta sección veremos cómo representar curvas generales al usar un parámetro para expresar tanto las variables x como y. Cualquier buen libro sobre gráficas computacionales mostrará cómo es posible ampliar esta técnica para representar curvas generales y superficies en el espacio (consulte, por ejemplo, [FVFH].)

Figura 3.15



Una técnica paramétrica sencilla para determinar un polinomio o un polinomio por tramos para conectar los puntos $(x_0, y_0), (x_1, y_1), \ldots, (x_n, y_n)$ en el orden provisto consiste en usar un parámetro t en un intervalo $[t_0, t_n]$, con $t_0 < t_1 < \cdots < t_n$ y construir funciones de aproximación con

$$x_i = x(t_i) \quad \text{y} \quad y_i = y(t_i), \quad \text{para cada } i = 0, 1, \ldots, n.$$

El siguiente ejemplo muestra la técnica en el caso en que ambas funciones de aproximación son polinomios de interpolación de Lagrange.

Ejemplo 1 Construya un par de polinomios de Lagrange para aproximar la curva que se muestra en la figura 3.15, usando los puntos de datos en la curva.

Solución Existe flexibilidad al seleccionar el parámetro y nosotros elegiremos los puntos $\{t_i\}_{i=0}^4$ igualmente espaciados en [0, 1], lo cual provee los datos en la tabla 3.20.

Tabla 3.20

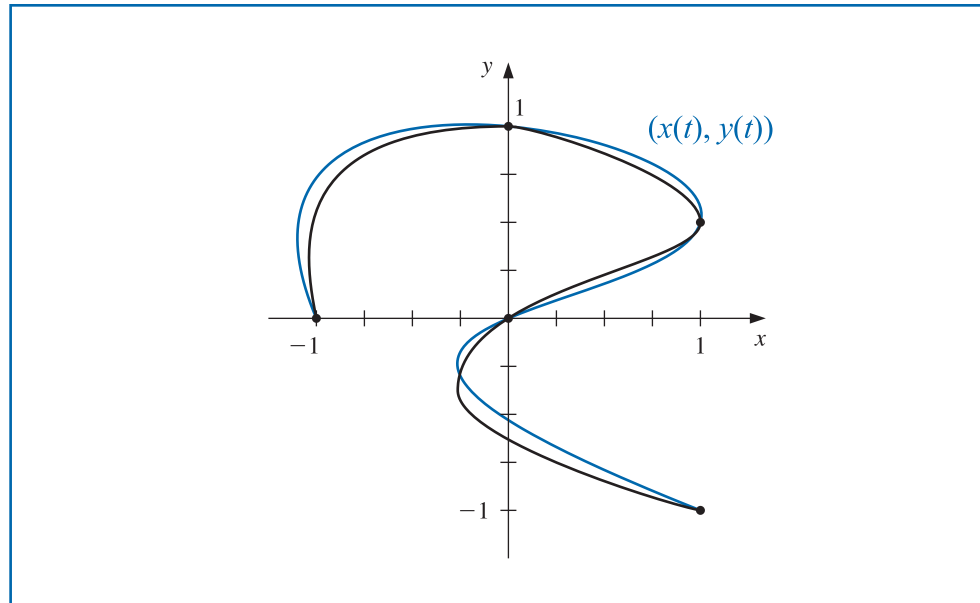
i	0	1	2	3	4
t_i	0	0.25	0.5	0.75	1
x_i	−1	0	1	0	1
y_i	0	1	0.5	0	−1

Esto produce los polinomios de interpolación

$$x(t) = \left(\left(\left(64t - \tfrac{352}{3}\right)t + 60\right)t - \tfrac{14}{3}\right)t - 1 \quad \text{y} \quad y(t) = \left(\left(\left(-\tfrac{64}{3}t + 48\right)t - \tfrac{116}{3}\right)t + 11\right)t.$$

Trazar este sistema paramétrico produce la gráfica que se muestra en azul en la figura 3.16. Aunque pasa por los puntos requeridos y tiene la misma forma básica, es una aproximación bastante burda para la curva original. Una aproximación más precisa requeriría nodos adicionales, con el consiguiente incremento en computación. ■

Figura 3.16



Las curvas paramétrica de Hermite y de spline se pueden generar de forma similar, pero esto también requiere un gran esfuerzo computacional.

Las aplicaciones de gráficas computacionales requieren la generación rápida de curvas suaves que se pueden modificar de manera fácil y rápida. Por razones tanto estéticas como computacionales, cambiar una parte de estas curvas debería tener un efecto pequeño o ningún efecto en otras partes de las curvas. Esto elimina el uso de polinomios de interpolación y splines ya que cambiar una parte de estas curvas afecta su totalidad.

Un sistema de diseño computacional exitoso necesita estar basado en una teoría matemática formal, de tal manera que los resultados sean predecibles, pero esta teoría debería realizarse en segundo plano para que el artista pueda basar el diseño en la estética.

La selección de la curva para usarla en gráficas computacionales es, en general, una forma de polinomio Hermite cúbico por tramos. Cada parte de un polinomio de Hermite cúbico se completa totalmente al especificar sus extremos y las derivadas en estos extremos. Por consiguiente, una parte de la curva se puede cambiar mientras la mayor parte de la misma se deja igual. Sólo las partes adyacentes necesitan modificarse para garantizar la suavidad en los extremos. Los cálculos se pueden realizar rápidamente y es posible cambiar una sección de la curva a la vez.

El problema con la interpolación de Hermite es la necesidad de especificar las derivadas en los extremos de cada sección de la curva. Suponga que la curva tiene $n + 1$ puntos de datos $(x(t_0), y(t_0)), \ldots, (x(t_n), y(t_n))$ y deseamos parametrizar el cúbico para permitir características complejas. Entonces, debemos especificar $x'(t_i)$ y $y'(t_i)$, para cada $i = 0, 1, \ldots, n$. Esto no es tan difícil como parece ya que cada parte se genera de manera independiente. Sólo debemos garantizar que las derivadas en los extremos de cada parte coincidan con los de la parte adyacente. En esencia, entonces, podemos simplificar el proceso en uno que determine un par de polinomios de Hermite cúbicos en el parámetro t, donde $t_0 = 0$ y $t_1 = 1$, dados los datos del extremo $(x(0), y(0))$ y $(x(1), y(1))$ y las derivadas dy/dx (en $t = 0$) y dy/dx (en $t = 1$).

Sin embargo, observe que sólo especificamos seis condiciones y los polinomios cúbicos en $x(t)$ y $y(t)$, cada uno tiene cuatro parámetros, para un total de ocho. Esto proporciona flexibilidad al seleccionar el par cúbico de polinomios de Hermite para satisfacer las condiciones porque la forma natural para determinar $x(t)$ y $y(t)$ requiere que especifiquemos $x'(0), x'(1), y'(0)$, y $y'(1)$. La curva de Hermite explícita en x y y requiere especificar solamente los cocientes

$$\frac{dy}{dx}(t=0) = \frac{y'(0)}{x'(0)} \quad \text{y} \quad \frac{dy}{dx}(t=1) = \frac{y'(1)}{x'(1)}.$$

Al multiplicar $x'(0)$ y $y'(0)$ por un factor de escala común, la recta tangente para la curva en $(x(0), y(0))$ permanece igual, pero la forma de la curva varía. Mientras más grande sea el factor de escala, más cerca está la curva de aproximación a la recta tangente en las cercanías de $(x(0), y(0))$. Existe una situación similar en el otro extremo $(x(1), y(1))$.

Para simplificar más el proceso en las gráficas computacionales interactivas, la derivada en un punto extremo se especifica usando un segundo punto, llamado *punto guía*, en una recta tangente deseada. Mientras más lejos esté del nodo, más cerca se aproxima la curva a la recta tangente cerca del nodo.

En la figura 3.17 los nodos se presentan en (x_0, y_0) y (x_1, y_1), el punto guía para (x_0, y_0) es $(x_0 + \alpha_0, y_0 + \beta_0)$, y el punto guía para (x_1, y_1) es $(x_1 - \alpha_1, y_1 - \beta_1)$. El polinomio de Hermite cúbico $x(t)$ en $[0, 1]$ satisface

$$x(0) = x_0, \quad x(1) = x_1, \quad x'(0) = \alpha_0, \quad \text{y} \quad x'(1) = \alpha_1.$$

El único polinomio cúbico que satisface estas condiciones es

$$x(t) = [2(x_0 - x_1) + (\alpha_0 + \alpha_1)]t^3 + [3(x_1 - x_0) - (\alpha_1 + 2\alpha_0)]t^2 + \alpha_0 t + x_0. \tag{3.23}$$

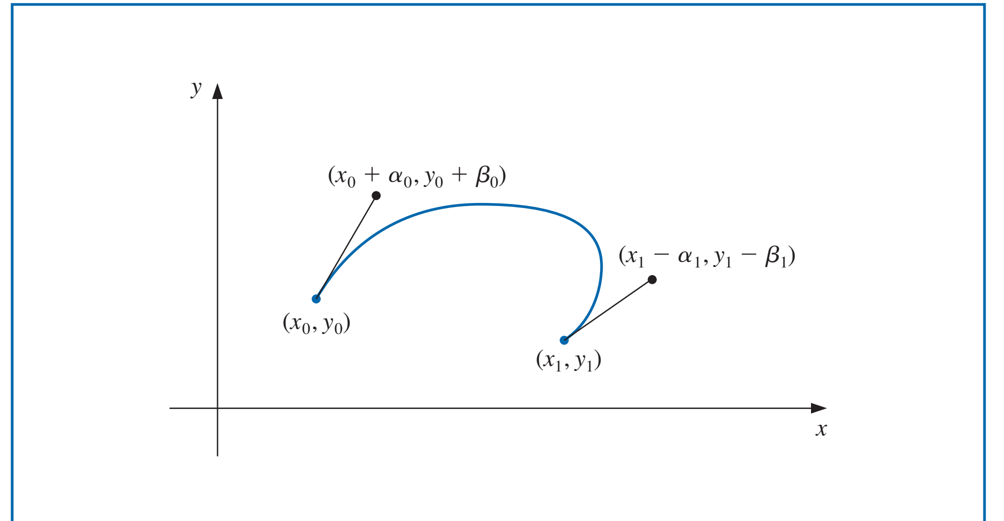
De manera similar, el único polinomio cúbico que satisface

$$y(0) = y_0, \quad y(1) = y_1, \quad y'(0) = \beta_0, \quad \text{y} \quad y'(1) = \beta_1$$

es

$$y(t) = [2(y_0 - y_1) + (\beta_0 + \beta_1)]t^3 + [3(y_1 - y_0) - (\beta_1 + 2\beta_0)]t^2 + \beta_0 t + y_0. \tag{3.24}$$

Figura 3.17



Ejemplo 2 Determine la gráfica de la curva paramétrica generada por las ecuaciones (3.23) y (3.24) cuando los extremos son $(x_0, y_0) = (0, 0)$ y $(x_1, y_1) = (1, 0)$ y los puntos guía respectivos, como se muestra en la figura 3.18 son (1, 1) y (0, 1).

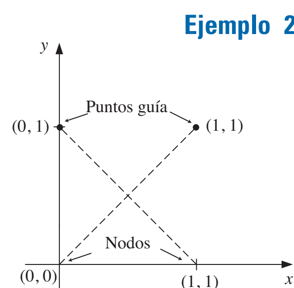


Figura 3.18

Pierre Etienne Bézier (1910–1999) fue director de diseño y producción de los automóviles Renault durante la mayor parte de su vida profesional. Comenzó su investigación en diseño y fabricación asistidos por computadora en 1960, desarrollando herramientas interactivas para el diseño de curvas y superficies, e inició el bobinado generado por computadora para modelado de automóviles. Las curvas de Bézier que llevan su nombre tienen la ventaja de estar basadas en una teoría matemática rigurosa que no necesita ser reconocida explícitamente por el practicante, quien sólo quiere crear una curva o superficie agradable desde el punto de vista estético. Éstas son las curvas que son la base del poderoso sistema Adobe Postscript y producen curvas a mano alzada generadas en la mayoría de los paquetes gráficos computacionales suficientemente potentes.

Solución La información del extremo implica que $x_0 = 0, x_1 = 1, y_0 = 0$, y $y_1 = 0$, y los puntos guía (1, 1) y (0, 1) implican que $\alpha_0 = 1, \alpha_1 = 1, \beta_0 = 1$, y $\beta_1 = -1$. Observe que las pendientes de las rectas guía en (0, 0) y (1, 0) son, respectivamente,

$$\frac{\beta_0}{\alpha_0} = \frac{1}{1} = 1 \quad \text{y} \quad \frac{\beta_1}{\alpha_1} = \frac{-1}{1} = -1.$$

Las ecuaciones (3.23) y (3.24) implican que para $t \in [0, 1]$, tenemos

$$x(t) = [2(0-1) + (1+1)]t^3 + [3(0-0) - (1+2\cdot 1)]t^2 + 1\cdot t + 0 = t$$

y

$$y(t) = [2(0-0) + (1+(-1))]t^3 + [3(0-0) - (-1+2\cdot 1)]t^2 + 1\cdot t + 0 = -t^2 + t.$$

Esta gráfica se muestra como a) en la figura 3.19, junto con algunas otras posibilidades de curvas producidas por las ecuaciones (3.23) y (3.24) cuando los nodos son (0, 0) y (1, 0) y las pendientes de estos nodos son 1 y −1, respectivamente. ■

El procedimiento estándar para determinar las curvas en un modo gráfico interactivo es utilizar primero un mouse (o ratón) o touchpad (panel táctil) y establecer los nodos y los puntos guía para generar una primera aproximación a la curva. Esto se puede hacer de manera manual, pero muchos sistemas de gráficas permiten utilizar su dispositivo de entrada para trazar la curva en una pantalla a mano alzada y seleccionar los nodos y puntos guía adecuados para su curva a mano alzada.

A continuación, los nodos y los puntos guía se pueden manipular en una posición que genera una curva agradable desde el punto de vista estético. Puesto que los cálculos son mínimos, la curva se puede determinar tan rápido que el cambio resultante se aprecia de inmediato. Además, todos los datos necesarios para calcular las curvas están incorporados en las coordenadas de los nodos y puntos guía, por lo que no se requiere que el usuario tenga conocimiento analítico.

Figura 3.19

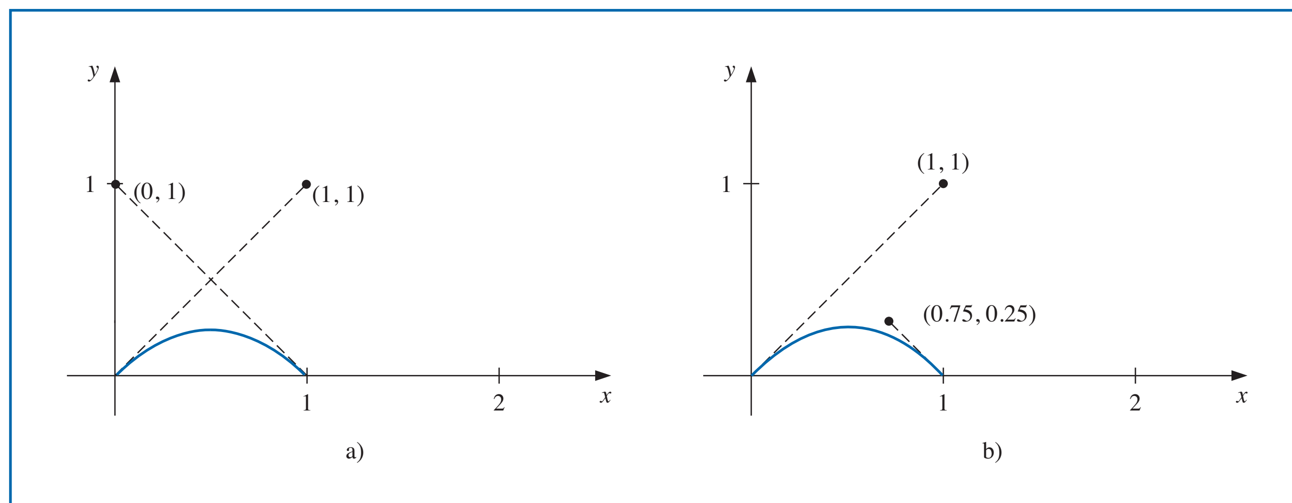
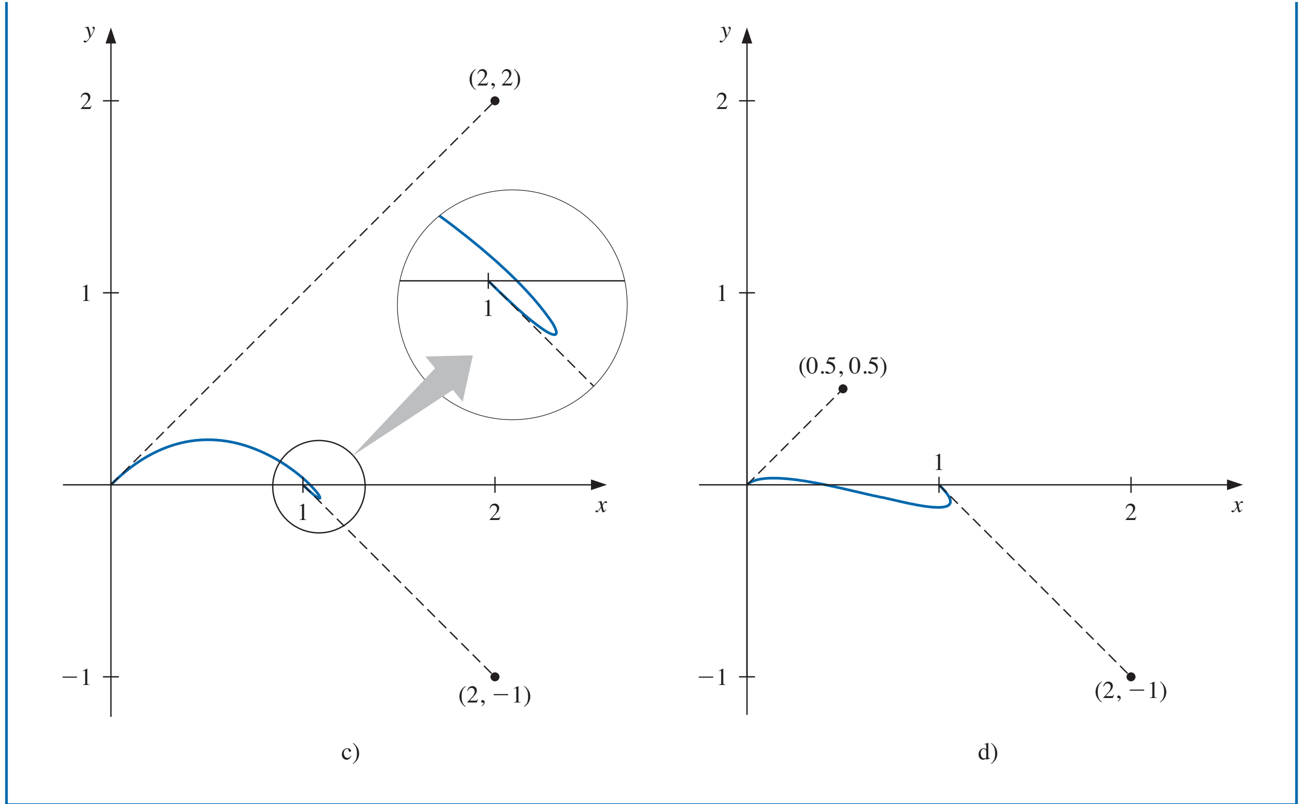


Figura 3.19



Los programas de gráficos populares usan este tipo de sistema para sus representaciones gráficas a mano alzada en una forma ligeramente modificada. Los cúbicos de Hermite se describen como **polinomios de Bézier**, los cuales incluyen un factor de escala de tres al calcular las derivadas en los extremos. Esto modifica las ecuaciones paramétricas para

$$x(t) = [2(x_0 - x_1) + 3(\alpha_0 + \alpha_1)]t^3 + [3(x_1 - x_0) - 3(\alpha_1 + 2\alpha_0)]t^2 + 3\alpha_0 t + x_0 \quad (3.25)$$

y

$$y(t) = [2(y_0 - y_1) + 3(\beta_0 + \beta_1)]t^3 + [3(y_1 - y_0) - 3(\beta_1 + 2\beta_0)]t^2 + 3\beta_0 t + y_0, \quad (3.26)$$

para $0 \leq t \leq 1$, pero este cambio es transparente para el usuario del sistema.

El algoritmo 3.6 construye un conjunto de curvas de Bézier con base en las ecuaciones paramétricas en las ecuaciones (3.25) y (3.26).

ALGORITMO 3.6

Curva de Bézier

Para construir las curvas cúbicas de Bézier $C_0, \ldots, C_{n-1}$ en forma paramétrica, donde C_i está representado por

$$(x_i(t), y_i(t)) = (a_0^{(i)} + a_1^{(i)}t + a_2^{(i)}t^2 + a_3^{(i)}t^3, b_0^{(i)} + b_1^{(i)}t + b_2^{(i)}t^2 + b_3^{(i)}t^3),$$

para $0 \leq t \leq 1$, como se determina mediante el extremo izquierdo (x_i, y_i), el punto guía (x_i^+, y_i^+), el extremo derecho (x_{i+1}, y_{i+1}), y el punto guía derecho (x_{i+1}^-, y_{i+1}^-) para cada $i = 0, 1, \ldots, n-1$:

ENTRADA n; $(x_0, y_0), \ldots, (x_n, y_n)$; $(x_0^+, y_0^+), \ldots, (x_{n-1}^+, y_{n-1}^+)$; $(x_1^-, y_1^-), \ldots, (x_n^-, y_n^-)$.

SALIDA coeficientes $\{a_0^{(i)}, a_1^{(i)}, a_2^{(i)}, a_3^{(i)}, b_0^{(i)}, b_1^{(i)}, b_2^{(i)}, b_3^{(i)},$ para $0 \leq i \leq n-1\}$.

Paso 1 Para cada $i = 0, 1, \ldots, n-1$ haga los pasos 2 y 3.

Paso 2 Haga $a_0^{(i)} = x_i$;

$b_0^{(i)} = y_i$;

$a_1^{(i)} = 3(x_i^+ - x_i)$;

$b_1^{(i)} = 3(y_i^+ - y_i)$;

$a_2^{(i)} = 3(x_i + x_{i+1}^- - 2x_i^+)$;

$b_2^{(i)} = 3(y_i + y_{i+1}^- - 2y_i^+)$;

$a_3^{(i)} = x_{i+1} - x_i + 3x_i^+ - 3x_{i+1}^-$;

$b_3^{(i)} = y_{i+1} - y_i + 3y_i^+ - 3y_{i+1}^-$;

Paso 3 SALIDA $(a_0^{(i)}, a_1^{(i)}, a_2^{(i)}, a_3^{(i)}, b_0^{(i)}, b_1^{(i)}, b_2^{(i)}, b_3^{(i)})$.

Paso 4 PARE. ■

Las curvas tridimensionales se generan de la misma forma al especificar adicionalmente terceros componentes z_0 y z_1 para los nodos y $z_0+\gamma_0$ y $z_1-\gamma_1$ para los puntos guía. El problema más difícil que implica la representación de las curvas tridimensionales se preocupa por la pérdida de la tercera dimensión cuando la curva se proyecta en una pantalla bidimensional de computadora. Se utilizan varias técnicas de proyección, pero este tema se encuentra dentro del campo de las gráficas por computador. Para una introducción a este tema y las maneras en las que la técnica se puede modificar para las representaciones de superficie, consulte alguno de los muchos libros sobre métodos gráficos por computadora como [FVFH].

La sección Conjunto de ejercicios 3.6 está disponible en línea. Encuentre la ruta de acceso en las páginas preliminares.

3.7 Software numérico y revisión del capítulo

Las rutinas de interpolación incluidas en la Biblioteca IMSL se basan en el libro *A practical Guide to Splines (Una guía práctica para splines)* de Carl de Boor [Deb] y usan la interpolación mediante splines cúbicos. Existen splines cúbicos para minimizar las oscilaciones y preservar la concavidad. Los métodos para interpolación bidimensional mediante splines bicúbicos también se incluyen.

La biblioteca NAG contiene subrutinas para la interpolación polinomial y de Hermite, para interpolación de spline cúbico y para interpolación de Hermite cúbico por tramos. NAG también contiene subrutinas para funciones de interpolación de dos variables.

La biblioteca netlib contiene las subrutinas para calcular el spline cúbico con varias condiciones de extremo. Un paquete produce los coeficientes de diferencia dividida de Newton para un conjunto discreto de puntos de datos y existen diferentes rutinas para evaluar los polinomios por tramos de Hermite.

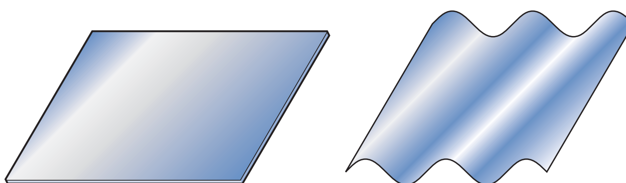
Las secciones Preguntas de análisis, Conceptos clave y Revisión del capítulo están disponibles en línea. Encuentre la ruta de acceso en las páginas preliminares.

CAPÍTULO

4 Diferenciación numérica e integración

Introducción

Una hoja de tejado corrugado se construye al presionar una hoja plana de aluminio dentro de otra cuya sección transversal tiene la forma de una onda senoidal.



Se necesita una hoja corrugada de 4 pies de largo, la altura de cada onda es de 1 pulgada desde la línea central y cada onda tiene un periodo de aproximadamente 2π. El problema de encontrar la longitud de la hoja plana inicial consiste en encontrar la longitud de la curva determinada por $f(x) = \text{sen}\, x$ desde $x = 0$ pulgadas hasta $x = 48$ pulgadas. A partir del cálculo, sabemos que esta longitud es

$$L = \int_0^{48} \sqrt{1 + (f'(x))^2}\, dx = \int_0^{48} \sqrt{1 + (\cos x)^2}\, dx,$$

por lo que el problema se reduce a evaluar esta integral. A pesar de que la función senoidal es una de las funciones matemáticas más comunes, el cálculo de su longitud implica una integral elíptica de segunda clase, que no se puede evaluar de manera explícita. En este capítulo se desarrollan métodos para aproximar la solución a los problemas de este tipo. Este problema particular se considera en el ejercicio 21 de la sección 4.4, en el ejercicio 15 de la sección 4.5 y en el ejercicio 10 de la sección 4.7.

En la introducción del capítulo 3 mencionamos que una razón para usar polinomios algebraicos para aproximar un conjunto arbitrario de datos es que, dada cualquier función continua definida dentro de un intervalo cerrado, existe un polinomio que está arbitrariamente cerca de la función en cada punto del intervalo. Además, las derivadas y las integrales de los polinomios se obtienen y evalúan con facilidad. No debería sorprender, entonces, que muchos procedimientos para aproximar derivadas e integrales utilicen los polinomios que aproximan la función.



4.1 Diferenciación numérica

La derivada de la función f en x_0 es

$$f'(x_0) = \lim_{h\to 0} \frac{f(x_0+h) - f(x_0)}{h}.$$

Esta fórmula proporciona una forma obvia de generar una aproximación para $f'(x_0)$; simplemente calcule

$$\frac{f(x_0+h) - f(x_0)}{h}$$

para valores pequeños de h. Aunque esto puede ser obvio, no tiene mucho éxito debido a nuestro antiguo némesis, el error de redondeo. Sin embargo, es un lugar para empezar.

Para aproximar $f'(x_0)$, suponga primero que $x_0 \in (a, b)$, donde $f \in C^2[a, b]$, y que $x_1 = x_0 + h$ para alguna $h \neq 0$, que es suficientemente pequeña para garantizar que $x_1 \in [a, b]$. Nosotros construimos el primer polinomio de Lagrange $P_{0,1}(x)$ para f determinado por x_0 y x_1, con su término de error:

$$\begin{aligned} f(x) &= P_{0,1}(x) + \frac{(x-x_0)(x-x_1)}{2!} f''(\xi(x)) \\ &= \frac{f(x_0)(x-x_0-h)}{-h} + \frac{f(x_0+h)(x-x_0)}{h} + \frac{(x-x_0)(x-x_0-h)}{2} f''(\xi(x)), \end{aligned}$$

para algunos $\xi(x)$ entre x_0 y x_1. Derivando obtenemos

$$\begin{aligned} f'(x) &= \frac{f(x_0+h) - f(x_0)}{h} + D_x \left[\frac{(x-x_0)(x-x_0-h)}{2} f''(\xi(x)) \right] \\ &= \frac{f(x_0+h) - f(x_0)}{h} + \frac{2(x-x_0) - h}{2} f''(\xi(x)) \\ &\quad + \frac{(x-x_0)(x-x_0-h)}{2} D_x(f''(\xi(x))). \end{aligned}$$

Borrando los términos relacionados con $\xi(x)$ obtenemos

$$f'(x) \approx \frac{f(x_0+h) - f(x_0)}{h}.$$

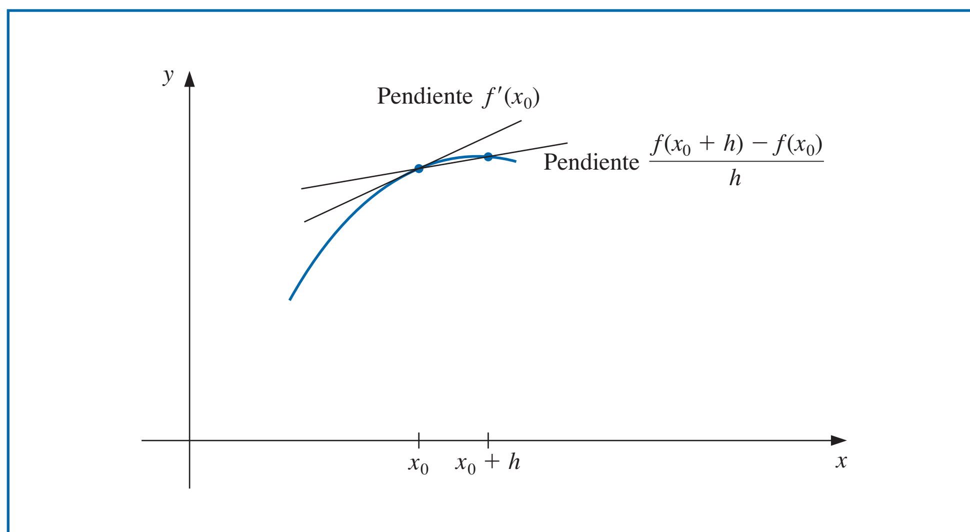
Una dificultad con esta fórmula es que no tenemos información sobre $D_x f''(\xi(x))$, por lo que el error de truncamiento no se puede calcular. Cuando x es x_0, sin embargo, el coeficiente de $D_x f''(\xi(x))$ es 0 y la fórmula se simplifica en

$$f'(x_0) = \frac{f(x_0+h) - f(x_0)}{h} - \frac{h}{2} f''(\xi). \tag{4.1}$$

Isaac Newton usó y popularizó las ecuaciones de diferencias en el último cuarto del siglo XVII, pero muchas de estas técnicas fueron desarrolladas previamente por Thomas Harriot (1561–1621) y Henry Briggs (1561–1630). Harriot realizó avances significativos en técnicas de navegación y Briggs fue la persona más responsable de la aceptación de logaritmos como auxiliares para el cálculo.

Para los valores pequeños de h, el cociente de diferencia $[f(x_0+h) - f(x_0)]/h$ se puede utilizar para aproximar $f'(x_0)$ con un error acotado por $M|h|/2$, donde M es una cota de $|f''(x)|$ para x entre x_0 y $x_0 + h$. A esta fórmula se le conoce como **fórmula de diferencias hacia adelante** si $h > 0$ (véase la figura 4.1) y como **fórmula de diferencias hacia atrás** si $h < 0$.

Figura 4.1



Ejemplo 1 Use la fórmula de diferencias hacia adelante para aproximar la derivada de $f(x) = \ln x$ en $x_0 = 1.8$ mediante $h = 0.1, h = 0.05$, y $h = 0.01$ y determine las cotas para los errores de aproximación.

Solución La fórmula de diferencias hacia adelante

$$\frac{f(1.8+h) - f(1.8)}{h}$$

con $h = 0.1$ nos da

$$\frac{\ln 1.9 - \ln 1.8}{0.1} = \frac{0.64185389 - 0.58778667}{0.1} = 0.5406722.$$

Puesto que $f''(x) = -1/x^2$ y $1.8 < \xi < 1.9$, una cota para este error de aproximación es

$$\frac{|hf''(\xi)|}{2} = \frac{|h|}{2\xi^2} < \frac{0.1}{2(1.8)^2} = 0.0154321.$$

La aproximación y las cotas de error cuando $h = 0.05$ y $h = 0.01$ se encuentran de manera similar y los resultados se muestran en la tabla 4.1.

Tabla 4.1

h	$f(1.8+h)$	$\frac{f(1.8+h) - f(1.8)}{h}$	$\frac{\vert h\vert}{2(1.8)^2}$
0.1	0.64185389	0.5406722	0.0154321
0.05	0.61518564	0.5479795	0.0077160
0.01	0.59332685	0.5540180	0.0015432

Puesto que $f'(x) = 1/x$, el valor exacto de $f'(1.8)$ es $0.55\overline{5}$, y en este caso la cota del error está bastante cerca del verdadero error de aproximación. ■

Para obtener las fórmulas generales de aproximación a la derivada, suponga que $\{x_0, x_1, \ldots, x_n\}$ son $(n+1)$ números distintos en algún intervalo I y que $f \in C^{n+1}(I)$. A partir del teorema 3.3 en la página 83,

$$f(x) = \sum_{k=0}^{n} f(x_k)L_k(x) + \frac{(x-x_0)\cdots(x-x_n)}{(n+1)!} f^{(n+1)}(\xi(x)),$$

para algunos $\xi(x)$ en I, donde $L_k(x)$ denota el k-ésimo coeficiente del polinomio de Lagrange para f en $x_0, x_1, \ldots, x_n$. Al derivar esta ecuación obtenemos

$$f'(x) = \sum_{k=0}^{n} f(x_k)L_k'(x) + D_x\left[\frac{(x-x_0)\cdots(x-x_n)}{(n+1!)}\right] f^{(n+1)}(\xi(x))$$
$$+ \frac{(x-x_0)\cdots(x-x_n)}{(n+1)!} D_x[f^{(n+1)}(\xi(x))].$$

De nuevo tenemos un problema al calcular el error de truncamiento a menos que x sea uno de los números x_j. En este caso, el término que multiplica $D_x[f^{(n+1)}(\xi(x))]$ es 0 y la fórmula se vuelve

$$f'(x_j) = \sum_{k=0}^{n} f(x_k)L_k'(x_j) + \frac{f^{(n+1)}(\xi(x_j))}{(n+1)!}\prod_{\substack{k=0\\k\neq j}}^{n}(x_j - x_k), \tag{4.2}$$

que recibe el nombre de **fórmula de (n + 1) puntos** para aproximar $f'(x_j)$.

En general, el uso de más puntos de evaluación en la ecuación (4.2) produce mayor precisión, a pesar de que el número de evaluaciones funcionales y el crecimiento del error de redondeo disuade un poco esto. Las fórmulas más comunes son las relacionadas con tres y cinco puntos de evaluación.

Primero derivamos algunas fórmulas útiles de tres puntos y consideramos aspectos de sus errores. Puesto que

$$L_0(x) = \frac{(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)}, \quad \text{tenemos} \quad L_0'(x) = \frac{2x-x_1-x_2}{(x_0-x_1)(x_0-x_2)}.$$

De igual forma,

$$L_1'(x) = \frac{2x-x_0-x_2}{(x_1-x_0)(x_1-x_2)} \quad \text{y} \quad L_2'(x) = \frac{2x-x_0-x_1}{(x_2-x_0)(x_2-x_1)}.$$

Por lo tanto, a partir de la ecuación (4.2),

$$f'(x_j) = f(x_0)\left[\frac{2x_j-x_1-x_2}{(x_0-x_1)(x_0-x_2)}\right] + f(x_1)\left[\frac{2x_j-x_0-x_2}{(x_1-x_0)(x_1-x_2)}\right]$$
$$+ f(x_2)\left[\frac{2x_j-x_0-x_1}{(x_2-x_0)(x_2-x_1)}\right] + \frac{1}{6}f^{(3)}(\xi_j)\prod_{\substack{k=0\\k\neq j}}^{2}(x_j-x_k), \tag{4.3}$$

para cada $j = 0, 1, 2$, donde la notación ξ_j indica que este punto depende de x_j.

Fórmulas de tres puntos

Las fórmulas a partir de la ecuación (4.3) se vuelven especialmente útiles si los nodos están espaciados de manera uniforme, es decir, cuando

$$x_1 = x_0 + h \quad \text{y} \quad x_2 = x_0 + 2h, \quad \text{para algunas} \quad h \neq 0.$$

Supondremos nodos igualmente espaciados a lo largo del resto de esta sección.

Por medio de la ecuación (4.3) con $x_j = x_0$, $x_1 = x_0 + h$, y $x_2 = x_0 + 2h$ obtenemos

$$f'(x_0) = \frac{1}{h}\left[-\frac{3}{2}f(x_0) + 2f(x_1) - \frac{1}{2}f(x_2)\right] + \frac{h^2}{3}f^{(3)}(\xi_0).$$

Al hacer lo mismo para $x_j = x_1$ obtenemos

$$f'(x_1) = \frac{1}{h}\left[-\frac{1}{2}f(x_0) + \frac{1}{2}f(x_2)\right] - \frac{h^2}{6}f^{(3)}(\xi_1)$$

y, para $x_j = x_2$,

$$f'(x_2) = \frac{1}{h}\left[\frac{1}{2}f(x_0) - 2f(x_1) + \frac{3}{2}f(x_2)\right] + \frac{h^2}{3}f^{(3)}(\xi_2).$$

Puesto que $x_1 = x_0 + h$ y $x_2 = x_0 + 2h$, estas fórmulas también se pueden expresar como

$$f'(x_0) = \frac{1}{h}\left[-\frac{3}{2}f(x_0) + 2f(x_0+h) - \frac{1}{2}f(x_0+2h)\right] + \frac{h^2}{3}f^{(3)}(\xi_0),$$

$$f'(x_0+h) = \frac{1}{h}\left[-\frac{1}{2}f(x_0) + \frac{1}{2}f(x_0+2h)\right] - \frac{h^2}{6}f^{(3)}(\xi_1), \quad \text{y}$$

$$f'(x_0+2h) = \frac{1}{h}\left[\frac{1}{2}f(x_0) - 2f(x_0+h) + \frac{3}{2}f(x_0+2h)\right] + \frac{h^2}{3}f^{(3)}(\xi_2).$$

Por cuestiones de conveniencia, la sustitución de la variable x_0 por $x_0 + h$ se usa en medio de la ecuación para cambiar esta fórmula por una aproximación para $f'(x_0)$. Un cambio similar, x_0 para $x_0 + 2h$, se utiliza en la última ecuación. Esto nos da tres fórmulas para aproximar $f'(x_0)$

$$f'(x_0) = \frac{1}{2h}[-3f(x_0) + 4f(x_0+h) - f(x_0+2h)] + \frac{h^2}{3}f^{(3)}(\xi_0),$$

$$f'(x_0) = \frac{1}{2h}[-f(x_0-h) + f(x_0+h)] - \frac{h^2}{6}f^{(3)}(\xi_1), \quad \text{y}$$

$$f'(x_0) = \frac{1}{2h}[f(x_0-2h) - 4f(x_0-h) + 3f(x_0)] + \frac{h^2}{3}f^{(3)}(\xi_2).$$

Finalmente, observe que la última de estas ecuaciones se puede obtener a partir de la primera simplemente al reemplazar h por $-h$, de modo que en realidad sólo son dos fórmulas:

Fórmula del extremo de tres puntos

- $$f'(x_0) = \frac{1}{2h}[-3f(x_0) + 4f(x_0+h) - f(x_0+2h)] + \frac{h^2}{3}f^{(3)}(\xi_0) \tag{4.4}$$

  donde ξ_0 se encuentra entre x_0 y $x_0 + 2h$.

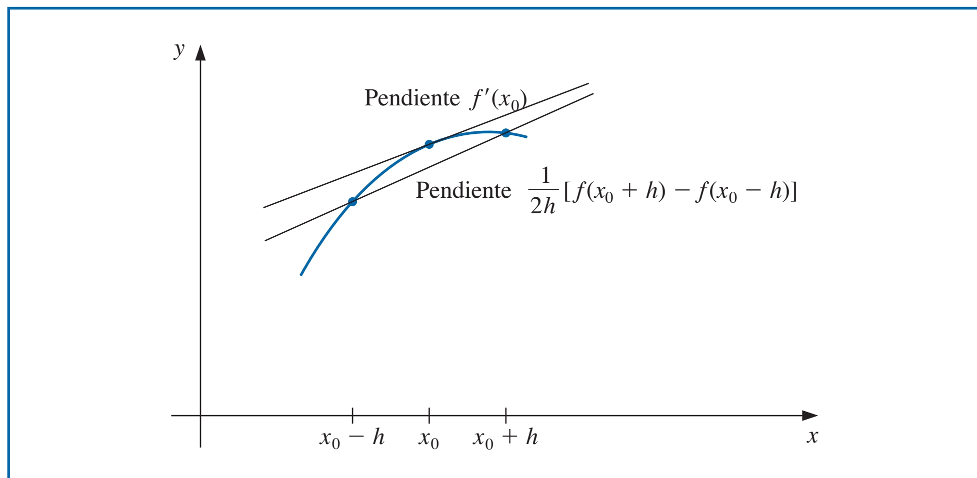
Fórmula del punto medio de tres puntos

- $$f'(x_0) = \frac{1}{2h}[f(x_0+h) - f(x_0-h)] - \frac{h^2}{6}f^{(3)}(\xi_1), \tag{4.5}$$

  donde ξ_1 se encuentra entre $x_0 - h$ y $x_0 + h$.

A pesar de que los errores en las dos ecuaciones (4.4) y (4.5) son $O(h^2)$, el error en la ecuación (4.5) es aproximadamente la mitad del error en la ecuación (4.4). Esto porque la ecuación (4.5) utiliza datos en ambos lados de x_0 y la ecuación (4.4) los usa en un solo lado. También observe que f se debe evaluar solamente en dos puntos en la ecuación (4.5), mientras que en la ecuación (4.4) se necesitan tres evaluaciones. La figura 4.2 ilustra la aproximación producida a partir de la ecuación (4.5). La aproximación en la ecuación (4.4) es útil cerca de los extremos de un intervalo porque la información sobre f fuera del intervalo puede no estar disponible.

Figura 4.2



Fórmulas de cinco puntos

Los métodos presentados en las ecuaciones (4.4) y (4.5) reciben el nombre de **fórmulas de tres puntos** (aunque el tercer punto $f(x_0)$ no aparece en la ecuación (4.5)). De igual forma, existen **fórmulas de cinco puntos** que implican la evaluación de la función en dos puntos adicionales. El término de error para estas fórmulas es $O(h^4)$. Una fórmula de cinco puntos común se usa para determinar las aproximaciones para la derivada en el punto medio.

Fórmula del punto medio de cinco puntos

- $$f'(x_0) = \frac{1}{12h}[f(x_0-2h) - 8f(x_0-h) + 8f(x_0+h) - f(x_0+2h)] + \frac{h^4}{30}f^{(5)}(\xi), \tag{4.6}$$

  donde ξ se encuentra entre $x_0 - 2h$ y $x_0 + 2h$.

La deducción de esta fórmula se considera en la sección 4.2. La otra fórmula de cinco puntos se usa para aproximaciones en los extremos.

Fórmula del extremo de cinco puntos

- $$f'(x_0) = \frac{1}{12h}[-25f(x_0) + 48f(x_0+h) - 36f(x_0+2h) + 16f(x_0+3h) - 3f(x_0+4h)] + \frac{h^4}{5}f^{(5)}(\xi), \tag{4.7}$$

  donde ξ se encuentra entre $x_0 + 4h$.

Las aproximaciones del extremo izquierdo se encuentran mediante esta fórmula con $h > 0$ y las aproximaciones del extremo derecho con $h < 0$. La fórmula del extremo de cinco puntos es especialmente útil para la interpolación de spline cúbico condicionado de la sección 3.5.

Ejemplo 2 Los valores para $f(x) = xe^x$ se dan en la tabla 4.2. Utilice las fórmulas aplicables de tres y cinco puntos para aproximar $f'(2.0)$.

Tabla 4.2

x	$f(x)$
1.8	10.889365
1.9	12.703199
2.0	14.778112
2.1	17.148957
2.2	19.855030

Solución Los datos en la tabla nos permiten encontrar cuatro diferentes aproximaciones de tres puntos. Podemos usar la fórmula del extremo (4.4) con $h = 0.1$ o con $h = -0.1$, y la fórmula del punto medio (4.5) con $h = 0.1$ o con $h = 0.2$.

Mediante la fórmula del extremo (4.4) con $h = 0.1$ obtenemos

$$\frac{1}{0.2}[-3f(2.0) + 4f(2.1) - f(2.2] = 5[-3(14.778112) + 4(17.148957) - 19.855030)]$$
$$= 22.032310$$

y con $h = -0.1$ nos da 22.054525.

Por medio de la fórmula del punto medio (4.5) con $h = 0.1$ obtenemos

$$\frac{1}{0.2}[f(2.1) - f(1.9] = 5(17.148957 - 12.7703199) = 22.228790$$

y con $h = 0.2$ nos da 22.414163.

La única fórmula de cinco puntos para la que la tabla provee datos suficientes es la fórmula del punto medio (4.6) con $h = 0.1$. Esto nos da

$$\frac{1}{1.2}[f(1.8) - 8f(1.9) + 8f(2.1) - f(2.2)] = \frac{1}{1.2}[10.889365 - 8(12.703199)$$
$$+ 8(17.148957) - 19.855030]$$
$$= 22.166999.$$

Si no tenemos más información, aceptaríamos la aproximación del punto medio de cinco puntos mediante $h = 0.1$ como la más apropiada y esperamos que el valor verdadero se encuentre entre esa aproximación y la aproximación del punto medio de tres puntos, es decir, en el intervalo [22.166, 22.229].

En este caso, el valor verdadero es $f'(2.0) = (2+1)e^2 = 22.167168$, por lo que, en realidad, los errores de aproximación son los siguientes:

Extremo de tres puntos con $h = 0.1$: 1.35×10^{-1};

Extremo de tres puntos con $h = -0.1$: 1.13×10^{-1};

Punto medio de tres puntos con $h = 0.1$: -6.16×10^{-2};

Punto medio de tres puntos con $h = 0.2$: -2.47×10^{-1};

Punto medio de cinco puntos con $h = 0.1$: 1.69×10^{-4}. ■

Los métodos también se pueden deducir para encontrar aproximaciones para derivadas superiores de una función que sólo usa valores tabulados de la función en diferentes puntos. La deducción es algebraicamente tediosa, sin embargo, sólo se presentará un procedimiento representativo.

Represente una función f mediante la expansión de un tercer polinomio de Taylor sobre un punto x_0 y evalúe $x_0 + h$ y $x_0 - h$. Entonces,

$$f(x_0 + h) = f(x_0) + f'(x_0)h + \frac{1}{2}f''(x_0)h^2 + \frac{1}{6}f'''(x_0)h^3 + \frac{1}{24}f^{(4)}(\xi_1)h^4$$

y

$$f(x_0 - h) = f(x_0) - f'(x_0)h + \frac{1}{2}f''(x_0)h^2 - \frac{1}{6}f'''(x_0)h^3 + \frac{1}{24}f^{(4)}(\xi_{-1})h^4,$$

donde $x_0 - h < \xi_{-1} < x_0 < \xi_1 < x_0 + h$.

Si adicionamos estas ecuaciones, los términos relacionados con $f'(x_0)$ y $-f'''(x_0)$ se cancelan, por lo que

$$f(x_0+h)+f(x_0-h)=2f(x_0)+f''(x_0)h^2+\frac{1}{24}[f^{(4)}(\xi_1)+f^{(4)}(\xi_{-1})]h^4.$$

Al resolver esta ecuación para $f''(x_0)$ obtenemos

$$f''(x_0)=\frac{1}{h^2}[f(x_0-h)-2f(x_0)+f(x_0+h)]-\frac{h^2}{24}[f^{(4)}(\xi_1)+f^{(4)}(\xi_{-1})]. \quad (4.8)$$

Suponga que $f^{(4)}$ es continua en $[x_0-h, x_0+h]$. Puesto que $\frac{1}{2}[f^{(4)}(\xi_1)+f^{(4)}(\xi_{-1})]$ se encuentra entre $f^{(4)}(\xi_1)$ y $f^{(4)}(\xi_{-1})$, el teorema de valor intermedio implica que existe un número ξ entre ξ_1 y ξ_{-1} y, por lo tanto, en (x_0-h, x_0+h) con

$$f^{(4)}(\xi)=\frac{1}{2}\left[f^{(4)}(\xi_1)+f^{(4)}(\xi_{-1})\right].$$

Esto nos permite reescribir la ecuación (4.8) en su forma final.

Fórmula del punto medio de la segunda derivada

- $$f''(x_0)=\frac{1}{h^2}[f(x_0-h)-2f(x_0)+f(x_0+h)]-\frac{h^2}{12}f^{(4)}(\xi), \quad (4.9)$$

  Para algunos ξ, donde $x_0-h<\xi<x_0+h$.

Si $f^{(4)}$ es continua en $[x_0-h, x_0+h]$, también está acotada, y la aproximación es $O(h^2)$.

Ejemplo 3 En el ejemplo 2 utilizamos los datos que se muestran en la tabla 4.3 para aproximar la primera derivada de $f(x)=xe^x$ en $x=2.0$. Use la fórmula de la segunda derivada (4.9) para aproximar $f''(2.0)$.

Tabla 4.3

x	$f(x)$
1.8	10.889365
1.9	12.703199
2.0	14.778112
2.1	17.148957
2.2	19.855030

Solución Los datos nos permiten determinar dos aproximaciones para $f''(2.0)$. Usando (4.9) con $h=0.1$ obtenemos

$$\frac{1}{0.01}[f(1.9)-2f(2.0)+f(2.1)]=100[12.703199-2(14.778112)+17.148957]$$
$$=29.593200,$$

y mediante (4.9) con $h=0.2$ obtenemos

$$\frac{1}{0.04}[f(1.8)-2f(2.0)+f(2.2)]=25[10.889365-2(14.778112)+19.855030]$$
$$=29.704275.$$

Puesto que $f''(x)=(x+2)e^x$, el valor exacto es $f''(2.0)=29.556224$. Por lo tanto, los errores reales son -3.70×10^{-2} y -1.48×10^{-1}, respectivamente. ■

Inestabilidad del error de redondeo

Es especialmente importante prestar atención al error de redondeo al aproximar derivadas. Para ilustrar la situación, examinemos más de cerca la fórmula del punto medio de tres puntos, la ecuación (4.5).

$$f'(x_0)=\frac{1}{2h}[f(x_0+h)-f(x_0-h)]-\frac{h^2}{6}f^{(3)}(\xi_1),$$

Suponga que al evaluar $f(x_0+h)$ y $f(x_0-h)$, encontramos los errores de redondeo $e(x_0+h)$ y $e(x_0-h)$. Entonces nuestros cálculos en realidad usan los valores $\tilde{f}(x_0+h)$ y $\tilde{f}(x_0-h)$, que están relacionados con los valores verdaderos $f(x_0+h)$ y $f(x_0-h)$ mediante

$$f(x_0+h) = \tilde{f}(x_0+h) + e(x_0+h) \quad \text{y} \quad f(x_0-h) = \tilde{f}(x_0-h) + e(x_0-h).$$

El error total en la aproximación,

$$f'(x_0) - \frac{\tilde{f}(x_0+h) - \tilde{f}(x_0-h)}{2h} = \frac{e(x_0+h) - e(x_0-h)}{2h} - \frac{h^2}{6} f^{(3)}(\xi_1),$$

se debe tanto al error de redondeo, la primera parte, como al error de truncamiento. Si suponemos que los errores de redondeo $e(x_0 \pm h)$ están acotados por algún número $\varepsilon > 0$ y que la tercera derivada de f está acotada por un número $M > 0$, entonces

$$\left| f'(x_0) - \frac{\tilde{f}(x_0+h) - \tilde{f}(x_0-h)}{2h} \right| \leq \frac{\varepsilon}{h} + \frac{h^2}{6} M.$$

Para reducir el error de truncamiento $h^2M/6$, necesitamos reducir h. Pero conforme h se reduce, el error de redondeo ε/h crece. En la práctica, entonces, casi nunca es ventajoso dejar que h sea demasiado pequeña, porque en este caso, el error de redondeo dominará los cálculos.

Ilustración Considere utilizar los valores en la tabla 4.4. Para aproximar $f'(0.900)$, donde $f(x) = \text{sen}\, x$. . El verdadero valor es $\cos 0.900 = 0.62161$. La fórmula

$$f'(0.900) \approx \frac{f(0.900+h) - f(0.900-h)}{2h},$$

con diferentes valores de h, proporciona las aproximaciones en la tabla 4.5.

Tabla 4.4

x	sen x	x	sen x
0.800	0.71736	0.901	0.78395
0.850	0.75128	0.902	0.78457
0.880	0.77074	0.905	0.78643
0.890	0.77707	0.910	0.78950
0.895	0.78021	0.920	0.79560
0.898	0.78208	0.950	0.81342
0.899	0.78270	1.000	0.84147

Tabla 4.5

h	Aproximación para $f'(0.900)$	Error
0.001	0.62500	0.00339
0.002	0.62250	0.00089
0.005	0.62200	0.00039
0.010	0.62150	−0.00011
0.020	0.62150	−0.00011
0.050	0.62140	−0.00021
0.100	0.62055	−0.00106

La mejor opción para h parece encontrarse entre 0.005 y 0.05. Podemos utilizar el cálculo para verificar (consulte el ejercicio 29) que se presenta un mínimo para

$$e(h) = \frac{\varepsilon}{h} + \frac{h^2}{6} M,$$

en $h = \sqrt[3]{3\varepsilon/M}$, donde

$$M = \max_{x \in [0.800, 1.00]} |f'''(x)| = \max_{x \in [0.800, 1.00]} |\cos x| = \cos 0.8 \approx 0.69671.$$

Puesto que los valores de f están determinados para cinco lugares decimales, supondremos que el error de redondeo está limitado por $\varepsilon = 5 \times 10^{-6}$. Por lo tanto, la mejor opción de h es aproximadamente

$$h = \sqrt[3]{\frac{3(0.000005)}{0.69671}} \approx 0.028,$$

que es consistente con los resultados en la tabla 4.6.

En la práctica no podemos calcular una h óptima para utilizarla en la aproximación de la derivada ya que no conocemos la tercera derivada de la función. Sin embargo, debemos seguir siendo conscientes de que reducir el tamaño del paso no siempre mejora la aproximación. ■

Sólo hemos considerado los problemas de error de redondeo presentados por la fórmula de tres puntos, ecuación (4.5), pero se presentan dificultades similares con todas las fórmulas de diferenciación. La razón se puede rastrear a la necesidad de dividir entre una potencia de h. Como encontramos en la sección 1.2 (consulte, especialmente, el ejemplo 3), la división entre números pequeños tiende a exagerar el error de redondeo y, de ser posible, debería evitarse esta operación. En caso de diferenciación numérica, no podemos evitar el problema por completo, a pesar de que los métodos de orden superior reducen la dificultad.

Tome en cuenta que las aproximaciones por método de diferencias pueden ser inestables.

Al igual que los métodos de aproximación, la diferenciación numérica es *inestable* ya que los valores pequeños de h necesarios para reducir el error de truncamiento también causan que el error de redondeo crezca. Esta es la primera clase de métodos inestables que hemos encontrado y, de ser posible, estas técnicas deberían evitarse. Sin embargo, además de usarse para propósitos computacionales, las fórmulas son necesarias para aproximar las soluciones de ecuaciones ordinarias y diferenciales parciales.

La sección Conjunto de ejercicios 4.1 está disponible en línea. Encuentre la ruta de acceso en las paginas preliminares

4.2 Extrapolación de Richardson

La extrapolación de Richardson se usa para generar resultados de alta precisión mientras usa fórmulas de bajo orden. A pesar de que el nombre adjunto al método se refiere a un artículo escrito por L. F. Richardson y J. A. Gaunt [RG] en 1927, la idea detrás de la técnica es mucho más antigua. Un artículo interesante respecto a la historia y la aplicación de la extrapolación se puede encontrar en [Joy].

La extrapolación se puede aplicar siempre que se sepa que una técnica de aproximación tiene un término de error con un formato predecible, uno que depende de un parámetro, normalmente el tamaño de paso h. Suponga que para cada número $h \neq 0$, tenemos una fórmula $N_1(h)$ que se aproxima a una constante M desconocida y que el error de truncamiento relacionado con la aproximación tiene la forma

$$M - N_1(h) = K_1 h + K_2 h^2 + K_3 h^3 + \cdots,$$

para algún conjunto de constantes $K_1, K_2, K_3, \ldots$. (desconocidas).

Lewis Fry Richardson (1881-1953) fue el primero en aplicar sistemáticamente las matemáticas a la predicción del tiempo mientras trabajaba en Inglaterra para la Oficina Meteorológica. Como opositor concienzudo durante la Primera Guerra Mundial, escribió ampliamente sobre la futilidad económica de la guerra, mediante sistemas de ecuaciones diferenciales para modelar interacciones racionales entre países. La técnica de extrapolación que lleva su nombre fue el redescubrimiento de una técnica con raíces que son tan antiguas como Christiaan Hugyens (1629–1695) y, posiblemente, Arquímedes (287–212 a.C.)

El error de truncamiento es $O(h)$, a menos que haya una variación más grande entre las constantes $K_1, K_2, K_3, \ldots$,

$$M - N_1(0.1) \approx 0.1 K_1, \quad M - N_1(0.01) \approx 0.01 K_1,$$

y, en general, $M - N_1(h) \approx K_1 h$.

El objetivo de la extrapolación es encontrar una forma fácil de combinarlas en lugar de aproximaciones $O(h)$ inapropiadas de manera adecuada para producir fórmulas con un error de truncamiento de orden superior.

Suponga, por ejemplo, que podemos combinar las fórmulas $N_1(h)$ para producir una fórmula de aproximación $O(h^2)$, $N_2(h)$, para M con

$$M - N_2(h) = \hat{K}_2 h^2 + \hat{K}_3 h^3 + \cdots,$$

para algún, nuevamente desconocido, conjunto de constantes $\hat{K}_2, \hat{K}_3, \ldots$. Entonces tenemos

$$M - N_2(0.1) \approx 0.01 \hat{K}_2, \quad M - N_2(0.01) \approx 0.0001 \hat{K}_2,$$

y así sucesivamente. Si las constantes K_1 y $\hat{K}_2$ son aproximadamente de la misma magnitud, entonces las aproximaciones $N_2(h)$ serían mucho mejores que las aproximaciones $N_1(h)$ correspondientes. La extrapolación continúa al combinar las aproximaciones $N_2(h)$ de manera que produce fórmulas con error de truncamiento $O(h^3)$ y así sucesivamente.

Para observar específicamente cómo podemos generar las fórmulas de extrapolación, considere la fórmula $O(h)$ para aproximar M

$$M = N_1(h) + K_1 h + K_2 h^2 + K_3 h^3 + \cdots . \tag{4.10}$$

Se asume que la fórmula se mantiene para todas las h positivas, por lo que reemplazamos el parámetro h a la mitad de su valor. A continuación, tenemos una segunda fórmula de aproximación $O(h)$

$$M = N_1\left(\frac{h}{2}\right) + K_1\frac{h}{2} + K_2\frac{h^2}{4} + K_3\frac{h^3}{8} + \cdots . \tag{4.11}$$

Restando dos veces la ecuación (4.11) de la ecuación (4.10) se elimina el término relacionado con K_1 y nos da

$$M = N_1\left(\frac{h}{2}\right) + \left[N_1\left(\frac{h}{2}\right) - N_2(h)\right] + K_2\left(\frac{h^2}{2} - h^2\right) + K_3\left(\frac{h^3}{4} - h^3\right) + \cdots . \tag{4.12}$$

Defina

$$N_2(h) = N_1\left(\frac{h}{2}\right) + \left[N_1\left(\frac{h}{2}\right) - N_1(h)\right].$$

Entonces, la ecuación (4.12) es una fórmula de aproximación $O(h^2)$ para M:

$$M = N_2(h) - \frac{K_2}{2}h^2 - \frac{3K_3}{4}h^3 - \cdots . \tag{4.13}$$

Ejemplo 1 En el ejemplo 1 de la sección 4.1, utilizamos el método de diferencias hacia adelante con $h = 0.1$ y $h = 0.05$ para encontrar aproximaciones para $f(1.8)$ para $f(x) = \ln(x)$. Suponga que esta fórmula tiene error de truncamiento $O(h)$. Use la extrapolación en estos valores para observar si esto resulta en una mejor aproximación.

Solución En el ejemplo 1 de la sección 4.1, encontramos que

$$\text{con } h = 0.1\text{: } f'(1.8) \approx 0.5406722, \quad \text{y} \quad \text{con } h = 0.05\text{: } f'(1.8) \approx 0.5479795.$$

Esto implica que

$$N_1(0.1) = 0.5406722 \quad \text{y} \quad N_1(0.05) = 0.5479795.$$

La extrapolación de estos resultados nos da la nueva aproximación

$$\begin{aligned} N_2(0.1) = N_1(0.05) + (N_1(0.05) - N_1(0.1)) &= 0.5479795 + (0.5479795 - 0.5406722) \\ &= 0.555287. \end{aligned}$$

Se encontró que los resultados $h = 0.1$ y $h = 0.05$ son precisos dentro de 1.5×10^{-2} y 7.7×10^{-3}, respectivamente. Puesto que $f'(1.8) = 1/1.8 = 0.\overline{5}$, el valor extrapolado es preciso dentro de 2.7×10^{-4}. ■

La extrapolación se puede aplicar siempre que el error de truncamiento para una fórmula tenga la forma

$$\sum_{j=1}^{m-1} K_j h^{\alpha_j} + O(h^{\alpha_m})$$

para un conjunto de constantes K_j y cuando $\alpha_1 < \alpha_2 < \alpha_3 < \cdots < \alpha_m$. Muchas fórmulas utilizadas en la extrapolación tienen errores de truncamiento que sólo contienen potencias pares de h, es decir, tienen la forma

$$M = N_1(h) + K_1 h^2 + K_2 h^4 + K_3 h^6 + \cdots . \tag{4.14}$$

La extrapolación es mucho más efectiva cuando todas las potencias de h están presentes debido a que el proceso de promediar genera resultados con errores $O(h^2), O(h^4), O(h^6), \ldots$, con esencialmente, ningún incremento en el cálculo, sobre los resultados con errores $O(h), O(h^2), O(h^3), \ldots$.

Suponga que la aproximación tiene la forma de la ecuación (4.14). Al reemplazar h con $h/2$ obtenemos la fórmula de aproximación $O(h^2)$

$$M = N_1\left(\frac{h}{2}\right) + K_1\frac{h^2}{4} + K_2\frac{h^4}{16} + K_3\frac{h^6}{64} + \cdots .$$

Al restar cuatro veces esta ecuación de la ecuación (4.14), se elimina el término h^2,

$$3M = \left[4N_1\left(\frac{h}{2}\right) - N_1(h)\right] + K_2\left(\frac{h^4}{4} - h^4\right) + K_3\left(\frac{h^6}{16} - h^6\right) + \cdots .$$

Dividiendo esta ecuación entre 3, produce una fórmula $O(h^4)$

$$M = \frac{1}{3}\left[4N_1\left(\frac{h}{2}\right) - N_1(h)\right] + \frac{K_2}{3}\left(\frac{h^4}{4} - h^4\right) + \frac{K_3}{3}\left(\frac{h^6}{16} - h^6\right) + \cdots .$$

Al definir

$$N_2(h) = \frac{1}{3}\left[4N_1\left(\frac{h}{2}\right) - N_1(h)\right] = N_1\left(\frac{h}{2}\right) + \frac{1}{3}\left[N_1\left(\frac{h}{2}\right) - N_1(h)\right]$$

produce la fórmula de la aproximación con error de truncamiento $O(h^4)$:

$$M = N_2(h) - K_2\frac{h^4}{4} - K_3\frac{5h^6}{16} + \cdots . \tag{4.15}$$

Ahora reemplace h en la ecuación (4.15) con $h/2$ para producir una segunda fórmula $O(h^4)$

$$M = N_2\left(\frac{h}{2}\right) - K_2\frac{h^4}{64} - K_3\frac{5h^6}{1024} - \cdots .$$

Al restar 16 veces esta ecuación de la ecuación (4.15) elimina el término h^4 y da

$$15M = \left[16N_2\left(\frac{h}{2}\right) - N_2(h)\right] + K_3\frac{15h^6}{64} + \cdots .$$

Al dividir esta ecuación entre 15 produce la nueva fórmula $O(h^6)$

$$M = \frac{1}{15}\left[16N_2\left(\frac{h}{2}\right) - N_2(h)\right] + K_3\frac{h^6}{64} + \cdots .$$

Ahora tenemos la fórmula de aproximación $O(h^6)$

$$N_3(h) = \frac{1}{15}\left[16N_2\left(\frac{h}{2}\right) - N_2(h)\right] = N_2\left(\frac{h}{2}\right) + \frac{1}{15}\left[N_2\left(\frac{h}{2}\right) - N_2(h)\right].$$

Al continuar con este procedimiento obtenemos, para cada $j = 2, 3, \ldots,$ la aproximación $O(h^{2j})$

$$N_j(h) = N_{j-1}\left(\frac{h}{2}\right) + \frac{N_{j-1}(h/2) - N_{j-1}(h)}{4^{j-1} - 1}.$$

La tabla 4.6 muestra el orden en el que se generan las aproximaciones cuando

$$M = N_1(h) + K_1h^2 + K_2h^4 + K_3h^6 + \cdots . \tag{4.16}$$

Se supone de manera conservadora que el verdadero resultado es preciso por lo menos dentro del acuerdo de los dos resultados inferiores en la diagonal, en este caso, dentro de $|N_3(h) - N_4(h)|$.

Tabla 4.6

$O(h^2)$	$O(h^4)$	$O(h^6)$	$O(h^8)$
1: $N_1(h)$			
2: $N_1(\frac{h}{2})$	**3:** $N_2(h)$		
4: $N_1(\frac{h}{4})$	**5:** $N_2(\frac{h}{2})$	**6:** $N_3(h)$	
7: $N_1(\frac{h}{8})$	**8:** $N_2(\frac{h}{4})$	**9:** $N_3(\frac{h}{2})$	**10:** $N_4(h)$

Ejemplo 2 El teorema de Taylor se puede utilizar para mostrar que la fórmula de diferencias centradas en la ecuación (4.5) para aproximar $f'(x_0)$ se puede expresar con una fórmula de error:

$$f'(x_0) = \frac{1}{2h}[f(x_0 + h) - f(x_0 - h)] - \frac{h^2}{6}f'''(x_0) - \frac{h^4}{120}f^{(5)}(x_0) - \cdots .$$

Encuentre las aproximaciones de orden $O(h^2)$, $O(h^4)$, y $O(h^6)$ para $f'(2.0)$ cuando $f = xe^x$ y $h = 0.2$.

Solución Probablemente, las constantes $K_1 = -f'''(x_0)/6$, $K_2 = -f^{(5)}(x_0)/120, \cdots,$ serán desconocidas, sin embargo, esto no es importante. Sólo necesitamos saber que estas constantes existen con el fin de aplicar la extrapolación.

Tenemos la aproximación $O(h^2)$

$$f'(x_0) = N_1(h) - \frac{h^2}{6}f'''(x_0) - \frac{h^4}{120}f^{(5)}(x_0) - \cdots , \tag{4.17}$$

donde

$$N_1(h) = \frac{1}{2h}[f(x_0 + h) - f(x_0 - h)].$$

Esto nos da las primeras aproximaciones $O(h^2)$

$$N_1(0.2) = \frac{1}{0.4}[f(2.2) - f(1.8)] = 2.5(19.855030 - 10.889365) = 22.414160$$

y

$$N_1(0.1) = \frac{1}{0.2}[f(2.1) - f(1.9)] = 5(17.148957 - 12.703199) = 22.228786.$$

Al combinarlas para producir la primera aproximación $O(h^4)$ obtenemos

$$N_2(0.2) = N_1(0.1) + \frac{1}{3}(N_1(0.1) - N_1(0.2)) = 22.228786 + \frac{1}{3}(22.228786 - 22.414160)$$
$$= 22.166995.$$

y

$$f(x_0 - h) = f(x_0) - f'(x_0)h + \frac{1}{2}f''(x_0)h^2 - \frac{1}{6}f'''(x_0)h^3$$
$$+ \frac{1}{24}f^{(4)}(x_0)h^4 - \frac{1}{120}f^{(5)}(\xi_2)h^5, \quad (4.19)$$

donde $x_0 - h < \xi_2 < x_0 < \xi_1 < x_0 + h$.

Al restar la ecuación (4.19) de la ecuación (4.18) obtenemos una nueva aproximación para $f'(x)$):

$$f(x_0 + h) - f(x_0 - h) = 2hf'(x_0) + \frac{h^3}{3}f'''(x_0) + \frac{h^5}{120}[f^{(5)}(\xi_1) + f^{(5)}(\xi_2)], \quad (4.20)$$

lo cual implica que

$$f'(x_0) = \frac{1}{2h}[f(x_0 + h) - f(x_0 - h)] - \frac{h^2}{6}f'''(x_0) - \frac{h^4}{240}[f^{(5)}(\xi_1) + f^{(5)}(\xi_2)].$$

Si $f^{(5)}$ es continua en $[x_0 - h, x_0 + h]$, el teorema del valor intermedio 1.11 implica que existe un número $\tilde{\xi}$ en $(x_0 - h, x_0 + h)$ con

$$f^{(5)}(\tilde{\xi}) = \frac{1}{2}\left[f^{(5)}(\xi_1) + f^{(5)}(\xi_2)\right].$$

Como consecuencia, tenemos la aproximación $O(h^2)$

$$f'(x_0) = \frac{1}{2h}[f(x_0 + h) - f(x_0 - h)] - \frac{h^2}{6}f'''(x_0) - \frac{h^4}{120}f^{(5)}(\tilde{\xi}). \quad (4.21)$$

A pesar de que la aproximación en la ecuación (4.21) es igual a la que se obtuvo en la fórmula de tres puntos en la ecuación (4.5), ahora, el punto desconocido de evaluación se presenta en $f^{(5)}$ en lugar de en f'''. La extrapolación aprovecha esto al reemplazar primero h en la ecuación (4.21) con $2h$ para producir la nueva fórmula

$$f'(x_0) = \frac{1}{4h}[f(x_0 + 2h) - f(x_0 - 2h)] - \frac{4h^2}{6}f'''(x_0) - \frac{16h^4}{120}f^{(5)}(\hat{\xi}), \quad (4.22)$$

donde $\hat{\xi}$ se encuentra entre $x_0 - 2h$ y $x_0 + 2h$.

Al multiplicar la ecuación (4.21) por 4 y restar la ecuación (4.22) produce

$$3f'(x_0) = \frac{2}{h}[f(x_0 + h) - f(x_0 - h)] - \frac{1}{4h}[f(x_0 + 2h) - f(x_0 - 2h)]$$
$$- \frac{h^4}{30}f^{(5)}(\tilde{\xi}) + \frac{2h^4}{15}f^{(5)}(\hat{\xi}).$$

Incluso si $f^{(5)}$ es continua en $[x_0 - 2h, x_0 + 2h]$, el teorema de valor intermedio 1.11 no se puede aplicar como lo hicimos para derivar la ecuación (4.21) porque tenemos la *diferencia* de términos relacionados con $f^{(5)}$. Sin embargo, es posible usar un método alterno para mostrar que $f^{(5)}(\tilde{\xi})$ y $f^{(5)}(\hat{\xi})$ se puede seguir reemplazando con un valor común $f^{(5)}(\xi)$. Al suponer esto y dividir entre 3 produce la fórmula del punto medio de cinco puntos la ecuación (4.6) que observamos en la sección 4.1:

$$f'(x_0) = \frac{1}{12h}[f(x_0 - 2h) - 8f(x_0 - h) + 8f(x_0 + h) - f(x_0 + 2h)] + \frac{h^4}{30}f^{(5)}(\xi). \quad ■$$

Otras fórmulas para las primeras derivadas y las derivadas superiores se pueden deducir de manera similar. Consulte, por ejemplo, el ejercicio 8.

A lo largo del texto se utiliza la técnica de extrapolación. Las aplicaciones más prominentes se presentan al aproximar las integrales en la sección 4.5 y al determinar soluciones aproximadas para las ecuaciones diferenciales en la sección 5.8.

La sección Conjunto de ejercicios 4.2 está disponible en línea. Encuentre la ruta de acceso en las páginas preliminares.

4.3 Elementos de integración numérica

A menudo surge la necesidad de evaluar la integral definida de una función que no tiene una antiderivada o cuya antiderivada no es fácil de obtener. El método básico asociado con la aproximación de $\int_a^b f(x)\,dx$ recibe el nombre de **cuadratura numérica**. Éste utiliza una suma $\sum_{i=0}^{n} a_i f(x_i)$ para aproximar $\int_a^b f(x)\,dx$.

Los métodos de cuadratura en esta sección se basan en los polinomios de interpolación que se han explicado en el capítulo 3. La idea básica es seleccionar un conjunto de nodos distintos $\{x_0, \ldots, x_n\}$ del intervalo $[a, b]$. Entonces integramos el polinomio interpolante de Lagrange

$$P_n(x) = \sum_{i=0}^{n} f(x_i) L_i(x)$$

y su término de error de truncamiento sobre $[a, b]$ para obtener

$$\begin{aligned}\int_a^b f(x)\,dx &= \int_a^b \sum_{i=0}^{n} f(x_i) L_i(x)\,dx + \int_a^b \prod_{i=0}^{n} (x - x_i) \frac{f^{(n+1)}(\xi(x))}{(n+1)!}\,dx \\ &= \sum_{i=0}^{n} a_i f(x_i) + \frac{1}{(n+1)!} \int_a^b \prod_{i=0}^{n} (x - x_i) f^{(n+1)}(\xi(x))\,dx,\end{aligned}$$

donde $\xi(x)$ se encuentra en $[a, b]$ para cada x y

$$a_i = \int_a^b L_i(x)\,dx, \quad \text{para cada } i = 0, 1, \ldots, n.$$

La fórmula de cuadratura es, por lo tanto,

$$\int_a^b f(x)\,dx \approx \sum_{i=0}^{n} a_i f(x_i),$$

con un error dado por

$$E(f) = \frac{1}{(n+1)!} \int_a^b \prod_{i=0}^{n} (x - x_i) f^{(n+1)}(\xi(x))\,dx.$$

Antes de analizar la situación general de las fórmulas de cuadratura, consideremos las fórmulas producidas mediante el uso del primer y del segundo polinomios de Lagrange con nodos igualmente espaciados. Esto da la **regla trapezoidal** y la **regla de Simpson**, las cuales se presentan generalmente en cursos de cálculo.

La regla trapezoidal

Para derivar la regla trapezoidal (o regla del trapecio) para aproximar $\int_a^b f(x)\,dx$, sean $x_0 = a, x_1 = b, h = b - a$ y utilice el polinomio de Lagrange

$$P_1(x) = \frac{(x - x_1)}{(x_0 - x_1)} f(x_0) + \frac{(x - x_0)}{(x_1 - x_0)} f(x_1).$$

Entonces

$$\int_a^b f(x)\,dx = \int_{x_0}^{x_1}\left[\frac{(x-x_1)}{(x_0-x_1)}f(x_0)+\frac{(x-x_0)}{(x_1-x_0)}f(x_1)\right]dx \\ +\frac{1}{2}\int_{x_0}^{x_1} f''(\xi(x))(x-x_0)(x-x_1)\,dx. \tag{4.23}$$

El producto $(x-x_0)(x-x_1)$ no cambia de signo en $[x_0, x_1]$, por lo que el teorema del valor promedio ponderado para integrales 1.13 se puede aplicar al término de error para obtener, para algunos ξ en (x_0, x_1),

$$\int_{x_0}^{x_1} f''(\xi(x))(x-x_0)(x-x_1)\,dx = f''(\xi)\int_{x_0}^{x_1}(x-x_0)(x-x_1)\,dx$$

$$= f''(\xi)\left[\frac{x^3}{3}-\frac{(x_1+x_0)}{2}x^2+x_0x_1x\right]_{x_0}^{x_1}$$

$$= -\frac{h^3}{6}f''(\xi).$$

Cuando usamos el término *trapezoidal*, nos referimos a una figura de cuatro lados con al menos dos lados paralelos. El término europeo para esta figura es *trapezium*. Para confundir más el tema, la palabra europea *trapezoidal* se refiere a una figura de cuatro lados sin ningún lado igual y la palabra estadounidense para este tipo de figura es *trapezium*.

Por consiguiente, la ecuación (4.23) implica que

$$\int_a^b f(x)\,dx = \left[\frac{(x-x_1)^2}{2(x_0-x_1)}f(x_0)+\frac{(x-x_0)^2}{2(x_1-x_0)}f(x_1)\right]_{x_0}^{x_1} - \frac{h^3}{12}f''(\xi)$$

$$= \frac{(x_1-x_0)}{2}[f(x_0)+f(x_1)] - \frac{h^3}{12}f''(\xi).$$

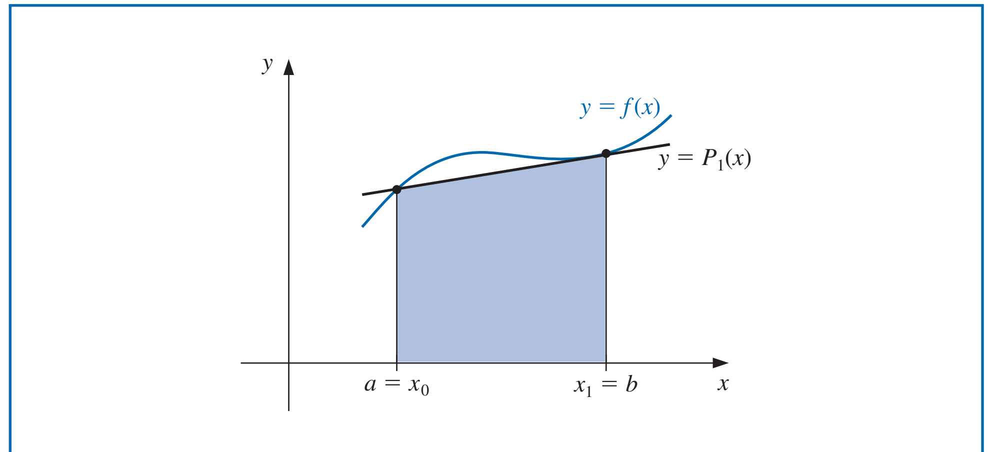
Por medio de la notación $h = x_1 - x_0$ obtenemos la siguiente regla:

Regla trapezoidal:

$$\int_a^b f(x)\,dx = \frac{h}{2}[f(x_0)+f(x_1)] - \frac{h^3}{12}f''(\xi).$$

Esto recibe el nombre de regla trapezoidal porque cuando f es una función con valores positivos, $\int_a^b f(x)\,dx$ se aproxima mediante el área de un trapecio, como se muestra en la figura 4.3

Figura 4.3

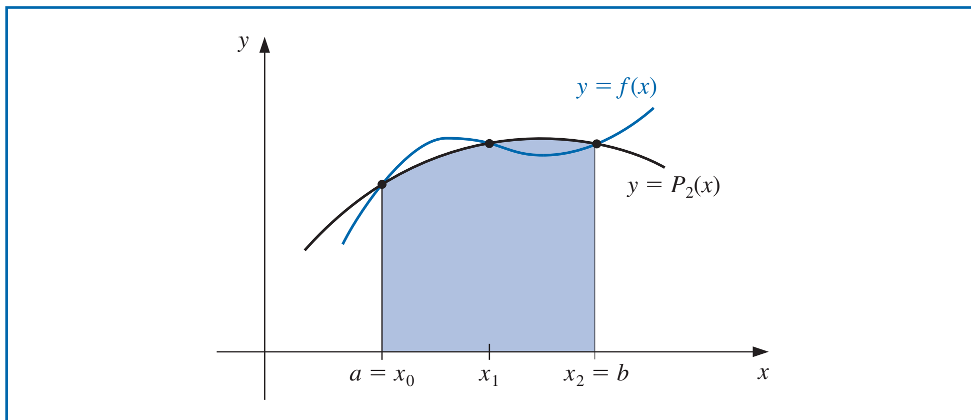


El término de error para la regla trapezoidal implica f'', por lo que la regla da el resultado exacto cuando se aplica a cualquier función cuya segunda derivada es idénticamente cero, es decir, cualquier polinomio de grado uno o menos.

Regla de Simpson

La regla de Simpson resulta de la integración sobre $[a, b]$ del segundo polinomio de Lagrange con nodos igualmente espaciados $x_0 = a$, $x_2 = b$, y $x_1 = a + h$, en donde $h = (b - a)/2$. (véase la figura 4.4).

Figura 4.4



Por lo tanto,

$$\int_a^b f(x)\,dx = \int_{x_0}^{x_2}\left[\frac{(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)}f(x_0) + \frac{(x-x_0)(x-x_2)}{(x_1-x_0)(x_1-x_2)}f(x_1)\right.$$
$$\left. + \frac{(x-x_0)(x-x_1)}{(x_2-x_0)(x_2-x_1)}f(x_2)\right]dx$$
$$+ \int_{x_0}^{x_2}\frac{(x-x_0)(x-x_1)(x-x_2)}{6}f^{(3)}(\xi(x))\,dx.$$

Al deducir la regla de Simpson de esta forma, sin embargo, da un solo término de error $O(h^4)$ relacionado con $f^{(3)}$. Al aproximar el problema de otra forma, se puede derivar otro término de orden superior relacionado con $f^{(4)}$.

Para ilustrar este método alternativo, suponga que f se expande en el tercer polinomio de Taylor alrededor de x_1. Entonces, para cada x en $[x_0, x_2]$, existe un número $\xi(x)$ en (x_0, x_2) con

$$f(x) = f(x_1) + f'(x_1)(x-x_1) + \frac{f''(x_1)}{2}(x-x_1)^2 + \frac{f'''(x_1)}{6}(x-x_1)^3$$
$$+ \frac{f^{(4)}(\xi(x))}{24}(x-x_1)^4$$

y

$$\int_{x_0}^{x_2} f(x)\,dx = \left[f(x_1)(x-x_1) + \frac{f'(x_1)}{2}(x-x_1)^2 + \frac{f''(x_1)}{6}(x-x_1)^3\right.$$
$$\left. + \frac{f'''(x_1)}{24}(x-x_1)^4\right]_{x_0}^{x_2} + \frac{1}{24}\int_{x_0}^{x_2} f^{(4)}(\xi(x))(x-x_1)^4\,dx. \quad (4.24)$$

Puesto que $(x - x_1)^4$ nunca es negativo en $[x_0, x_2]$, el teorema de valor promedio ponderado para las integrales 1.13 implica que

$$\frac{1}{24}\int_{x_0}^{x_2} f^{(4)}(\xi(x))(x-x_1)^4\,dx = \frac{f^{(4)}(\xi_1)}{24}\int_{x_0}^{x_2}(x-x_1)^4\,dx = \frac{f^{(4)}(\xi_1)}{120}(x-x_1)^5\bigg]_{x_0}^{x_2},$$

para algún número ξ_1 en (x_0, x_2).

Sin embargo, $h = x_2 - x_1 = x_1 - x_0$, por lo que

$$(x_2 - x_1)^2 - (x_0 - x_1)^2 = (x_2 - x_1)^4 - (x_0 - x_1)^4 = 0,$$

mientras

$$(x_2 - x_1)^3 - (x_0 - x_1)^3 = 2h^3 \quad \text{y} \quad (x_2 - x_1)^5 - (x_0 - x_1)^5 = 2h^5.$$

Por consiguiente, la ecuación (4.24) se puede reescribir como

$$\int_{x_0}^{x_2} f(x)\,dx = 2hf(x_1) + \frac{h^3}{3} f''(x_1) + \frac{f^{(4)}(\xi_1)}{60} h^5.$$

Ahora, si reemplazamos $f''(x_1)$ por medio de la aproximación determinada en la ecuación (4.9) de la sección 4.1, tenemos

$$\int_{x_0}^{x_2} f(x)\,dx = 2hf(x_1) + \frac{h^3}{3}\left\{\frac{1}{h^2}[f(x_0) - 2f(x_1) + f(x_2)] - \frac{h^2}{12} f^{(4)}(\xi_2)\right\} + \frac{f^{(4)}(\xi_1)}{60} h^5$$

$$= \frac{h}{3}[f(x_0) + 4f(x_1) + f(x_2)] - \frac{h^5}{12}\left[\frac{1}{3} f^{(4)}(\xi_2) - \frac{1}{5} f^{(4)}(\xi_1)\right].$$

Con métodos alternos se puede mostrar (consulte el ejercicio 26) que los valores ξ_1 y ξ_2 en esta expresión se pueden reemplazar mediante un valor común ξ en (x_0, x_2). Esto da la regla de Simpson.

Thomas Simpson (1710–1761) fue un matemático autodidacta que en sus primeros años se ganaba la vida como tejedor. Su principal interés fue la teoría de la probabilidad, aunque en 1750 publicó un libro de cálculo de dos volúmenes titulado *The Doctrine and Application of Fluxions (La doctrina y aplicación de fluxiones).*

Regla de Simpson:

$$\int_{x_0}^{x_2} f(x)\,dx = \frac{h}{3}[f(x_0) + 4f(x_1) + f(x_2)] - \frac{h^5}{90} f^{(4)}(\xi).$$

El término de error en la regla de Simpson implica la cuarta derivada de f, por lo que da resultados exactos cuando se aplica a cualquier polinomio de grado tres o menos.

Ejemplo 1 Compare las aproximaciones de la regla trapezoidal y de la regla de Simpson para $\int_0^2 f(x)\,dx$ cuando $f(x)$ es

a)	x^2	**b)**	x^4	**c)**	$(x+1)^{-1}$
d)	$\sqrt{1+x^2}$	**e)**	$\operatorname{sen} x$	**f)**	e^x

Solución En [0, 2], las reglas trapezoidal y de Simpson tiene las formas

$$\text{Trapezoidal:} \quad \int_0^2 f(x)\,dx \approx f(0) + f(2) \quad \text{y}$$

$$\text{De Simpson:} \quad \int_0^2 f(x)\,dx \approx \frac{1}{3}[f(0) + 4f(1) + f(2)].$$

Cuando $f(x) = x^2$, obtenemos

$$\text{Trapezoidal:} \quad \int_0^2 f(x)\,dx \approx 0^2 + 2^2 = 4 \quad \text{y}$$

$$\text{De Simpson:} \quad \int_0^2 f(x)\,dx \approx \frac{1}{3}[(0^2) + 4 \cdot 1^2 + 2^2] = \frac{8}{3}.$$

La aproximación a partir de la regla de Simpson es exacta porque su error de truncamiento implica $f^{(4)}$, lo cual es idénticamente 0 cuando $f(x) = x^2$.

Los resultados a los tres lugares para las funciones se resumen en la tabla 4.7. Observe que en cada instancia, la regla de Simpson es significativamente superior. ■

Tabla 4.7

	(a)	(b)	(c)	(d)	(e)	(f)
$f(x)$	x^2	x^4	$(x+1)^{-1}$	$\sqrt{1+x^2}$	sen x	e^x
Valor exacto	2.667	6.400	1.099	2.958	1.416	6.389
Trapezoidal	4.000	16.000	1.333	3.326	0.909	8.389
De Simpson	2.667	6.667	1.111	2.964	1.425	6.421

Precisión de medición

Las derivadas estándar de las fórmulas de error de cuadratura están basadas al determinar la clase de polinomios para los que estas fórmulas producen resultados exactos. La siguiente definición se utiliza para facilitar el análisis de esta derivada.

Definición 4.1 El **grado de precisión**, o **precisión**, de una fórmula de cuadratura es el mayor entero positivo n, de tal forma que la fórmula es exacta para x^k, para cada $k = 0, 1, \ldots, n$.

La precisión mejorada de la regla de Simpson sobre la regla trapezoidal se explica de manera intuitiva con el hecho de que la regla de Simpson incluye una evaluación de punto medio que proporciona mejor equilibrio para la aproximación.

La definición 4.1 implica que las reglas trapezoidal y de Simpson tienen grados de precisión uno y tres, respectivamente.

La integración y la sumatoria son operaciones lineales; es decir,

$$\int_a^b (\alpha f(x) + \beta g(x))\, dx = \alpha \int_a^b f(x)\, dx + \beta \int_a^b g(x)\, dx$$

y

$$\sum_{i=0}^{n} (\alpha f(x_i) + \beta g(x_i)) = \alpha \sum_{i=0}^{n} f(x_i) + \beta \sum_{i=0}^{n} g(x_i),$$

para cada par de funciones integrables f y g y cada par de constantes reales α y β. Esto implica (consulte el ejercicio 25) que

La terminología abierta y cerrada para los métodos implica que el método abierto sólo utiliza como nodos los puntos en el intervalo abierto (a, b) para aproximar $\int_a^b f(x)\,dx$. Los métodos cerrados incluyen los puntos a y b del intervalo cerrado $[a, b]$ como nodos.

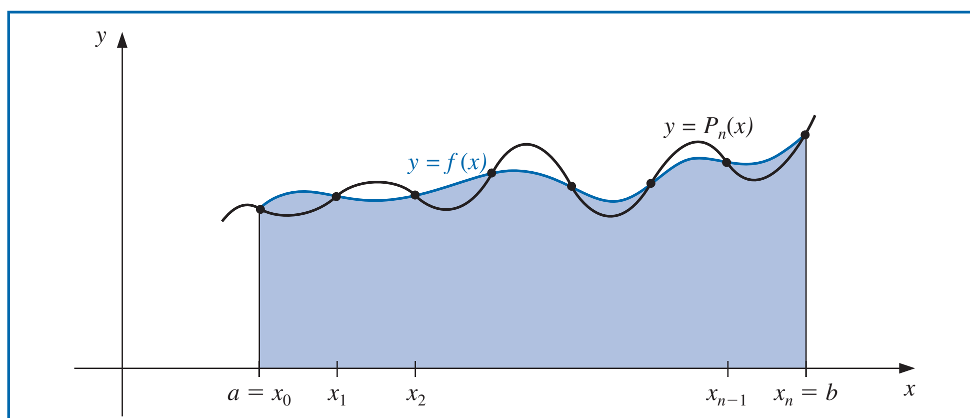
- el grado de precisión de una fórmula de cuadratura es n si y sólo si el error es cero para todos los polinomios de grado $k = 0, 1, \ldots, n$, pero no es cero para algunos polinomios de grado $n + 1$.

Las reglas trapezoidal y de Simpson son ejemplos de una clase de métodos conocidos como fórmulas de Newton-Cotes. Existen dos tipos de fórmulas de Newton-Cotes: abiertas y cerradas.

Fórmulas de Newton-Cotes cerradas

La *fórmula cerrada de* $(n + 1)$ *puntos de Newton-Cotes* utiliza nodos $x_i = x_0 + ih$, para $i = 0, 1, \ldots, n$, donde $x_0 = a$, $x_n = b$ y $h = (b - a)/n$. (Véase la figura 4.5.) Recibe el nombre de cerrada porque los extremos del intervalo cerrado $[a, b]$ se incluyen como nodos.

Figura 4.5



La fórmula asume la forma

$$\int_a^b f(x)\,dx \approx \sum_{i=0}^{n} a_i f(x_i),$$

donde

$$a_i = \int_{x_0}^{x_n} L_i(x)\,dx = \int_{x_0}^{x_n} \prod_{\substack{j=0\\ j\neq i}}^{n} \frac{(x - x_j)}{(x_i - x_j)}\,dx.$$

El siguiente teorema detalla el análisis de error relacionado con las fórmulas de Newton-Cotes cerradas. Para una demostración de este teorema, consulte [IK], p. 313.

Teorema 4.2 Suponga que $\sum_{i=0}^{n} a_i f(x_i)$ denota la fórmula cerrada de $(n+1)$ puntos de Newton-Cotes con $x_0 = a$, $x_n = b$, y $h = \ b - a)/n$. Existe $\xi \in (a, b)$ para el que

$$\int_a^b f(x)\,dx = \sum_{i=0}^{n} a_i f(x_i) + \frac{h^{n+3} f^{(n+2)}(\xi)}{(n+2)!} \int_0^n t^2(t-1)\cdots(t-n)\,dt,$$

si n es par y $f \in C^{n+2}[a, b]$, y

$$\int_a^b f(x)\,dx = \sum_{i=0}^{n} a_i f(x_i) + \frac{h^{n+2} f^{(n+1)}(\xi)}{(n+1)!} \int_0^n t(t-1)\cdots(t-n)\,dt,$$

si n es impar y $f \in C^{n+1}[a, b]$. ■

Roger Cotes (1682–1716) ascendió desde un origen humilde hasta convertirse, en 1704, en el primer profesor plumiano en la Universidad de Cambridge. Realizó numerosos avances en las áreas de matemáticas, incluyendo los métodos numéricos para la interpolación e integración. Newton es famoso por decir respecto a Cotes: “Si hubiera vivido, habríamos aprendido algo”.

Observe que cuando n es un entero par, el grado de precisión es $n + 1$, a pesar de que el polinomio de interpolación es de grado a lo sumo n. Cuando n es impar, el grado de precisión sólo es n.

Se listan algunas de las **fórmulas comunes de Newton-Cotes** cerradas con sus términos de error. Observe que, en cada caso, el valor desconocido ξ se encuentra en (a, b).

$n = 1$: Regla trapezoidal

$$\int_{x_0}^{x_1} f(x)\,dx = \frac{h}{2}[f(x_0) + f(x_1)] - \frac{h^3}{12} f''(\xi), \quad \text{donde} \quad x_0 < \xi < x_1. \tag{4.25}$$

$n = 2$: Regla de Simpson

$$\int_{x_0}^{x_2} f(x)\,dx = \frac{h}{3}[f(x_0) + 4f(x_1) + f(x_2)] - \frac{h^5}{90} f^{(4)}(\xi), \quad \text{donde} \quad x_0 < \xi < x_2. \tag{4.26}$$

$n = 3$: Regla de tres octavos de Simpson

$$\int_{x_0}^{x_3} f(x)\,dx = \frac{3h}{8}[f(x_0) + 3f(x_1) + 3f(x_2) + f(x_3)] - \frac{3h^5}{80} f^{(4)}(\xi), \tag{4.27}$$

$$\text{donde} \quad x_0 < \xi < x_3.$$

***n* = 4:**

$$\int_{x_0}^{x_4} f(x)\,dx = \frac{2h}{45}[7f(x_0) + 32f(x_1) + 12f(x_2) + 32f(x_3) + 7f(x_4)] - \frac{8h^7}{945}f^{(6)}(\xi),$$

$$\text{donde} \quad x_0 < \xi < x_4. \tag{4.28}$$

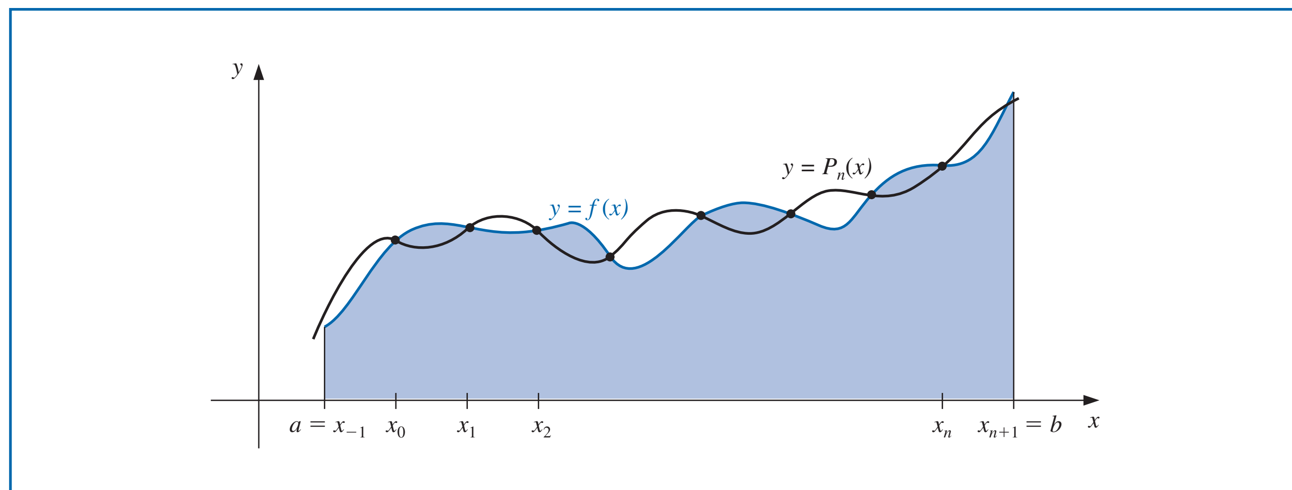
Fórmulas de Newton-Cotes abiertas

Las *fórmulas de Newton-Cotes abiertas* no incluyen los extremos de $[a, b]$ como nodos. Éstas utilizan los nodos $x_i = x_0 + ih$, para cada $i = 0, 1, \ldots, n$, donde $h = (b - a)/(n + 2)$ y $x_0 = a + h$. Esto implica que $x_n = b - h$, por lo que etiquetamos los extremos al establecer $x_{-1} = a$ y $x_{n+1} = b$, como se muestra en la figura 4.6. Las fórmulas abiertas contienen todos los nodos que se usan para la aproximación dentro del intervalo abierto (a, b). Las fórmulas se convertirán en

$$\int_a^b f(x)\,dx = \int_{x_{-1}}^{x_{n+1}} f(x)\,dx \approx \sum_{i=0}^{n} a_i f(x_i),$$

donde $a_i = \int_a^b L_i(x)\,dx$.

Figura 4.6



El siguiente teorema es análogo al teorema 4.2; su demostración se encuentra en [IK], p. 314.

Teorema 4.3 Suponga que $\sum_{i=0}^{n} a_i f(x_i)$ denota la fórmula abierta de $(n + 1)$ puntos de Newton-Cotes con $x_{-1} = a, x_{n+1} = b$, y $h = (b - a)/(n + 2)$. Existe $\xi \in (a, b)$ para el que

$$\int_a^b f(x)\,dx = \sum_{i=0}^{n} a_i f(x_i) + \frac{h^{n+3} f^{(n+2)}(\xi)}{(n+2)!} \int_{-1}^{n+1} t^2(t-1)\cdots(t-n)\,dt,$$

si n es par y $f \in C^{n+2}[a, b]$, y

$$\int_a^b f(x)\,dx = \sum_{i=0}^{n} a_i f(x_i) + \frac{h^{n+2} f^{(n+1)}(\xi)}{(n+1)!} \int_{-1}^{n+1} t(t-1)\cdots(t-n)\,dt,$$

si n es impar y $f \in C^{n+1}[a, b]$. ■

Observe que, como en el caso de métodos cerrados, tenemos el grado de precisión comparativamente superior para los métodos pares que para los métodos impares.

Algunas de las **fórmulas de Newton-Cotes abiertas** comunes con sus términos de error son las siguientes:

$n = 0$: regla del punto medio

$$\int_{x_{-1}}^{x_1} f(x)\,dx = 2hf(x_0) + \frac{h^3}{3} f''(\xi), \quad \text{donde} \quad x_{-1} < \xi < x_1. \tag{4.29}$$

$n = 1$:

$$\int_{x_{-1}}^{x_2} f(x)\,dx = \frac{3h}{2}[f(x_0) + f(x_1)] + \frac{3h^3}{4} f''(\xi), \quad \text{donde} \quad x_{-1} < \xi < x_2. \tag{4.30}$$

$n = 2$:

$$\int_{x_{-1}}^{x_3} f(x)\,dx = \frac{4h}{3}[2f(x_0) - f(x_1) + 2f(x_2)] + \frac{14h^5}{45} f^{(4)}(\xi), \tag{4.31}$$

$$\text{donde} \quad x_{-1} < \xi < x_3.$$

$n = 3$:

$$\int_{x_{-1}}^{x_4} f(x)\,dx = \frac{5h}{24}[11f(x_0) + f(x_1) + f(x_2) + 11f(x_3)] + \frac{95}{144} h^5 f^{(4)}(\xi), \tag{4.32}$$

$$\text{donde} \quad x_{-1} < \xi < x_4.$$

Ejemplo 2 Compare los resultados de las fórmulas cerradas y abiertas de Newton-Cotes como la ecuación (4.25) a la (4.28) y la ecuación (4.29) a la (4.32) para aproximar

$$\int_0^{\pi/4} \operatorname{sen} x\,dx = 1 - \sqrt{2}/2 \approx 0.29289322.$$

Solución Para las fórmulas cerradas, tenemos

$$n = 1: \quad \frac{(\pi/4)}{2}\left[\operatorname{sen} 0 + \operatorname{sen} \frac{\pi}{4}\right] \approx 0.27768018$$

$$n = 2: \quad \frac{(\pi/8)}{3}\left[\operatorname{sen} 0 + 4\operatorname{sen} \frac{\pi}{8} + \operatorname{sen} \frac{\pi}{4}\right] \approx 0.29293264$$

$$n = 3: \quad \frac{3(\pi/12)}{8}\left[\operatorname{sen} 0 + 3\operatorname{sen} \frac{\pi}{12} + 3\operatorname{sen} \frac{\pi}{6} + \operatorname{sen} \frac{\pi}{4}\right] \approx 0.29291070$$

$$n = 4: \quad \frac{2(\pi/16)}{45}\left[7\operatorname{sen} 0 + 32\operatorname{sen} \frac{\pi}{16} + 12\operatorname{sen} \frac{\pi}{8} + 32\operatorname{sen} \frac{3\pi}{16} + 7\operatorname{sen} \frac{\pi}{4}\right] \approx 0.29289318$$

y para las fórmulas abiertas, tenemos

$$n=0: \quad 2(\pi/8)\left[\operatorname{sen}\frac{\pi}{8}\right] \approx 0.30055887$$

$$n=1: \quad \frac{3(\pi/12)}{2}\left[\operatorname{sen}\frac{\pi}{12}+\operatorname{sen}\frac{\pi}{6}\right] \approx 0.29798754$$

$$n=2: \quad \frac{4(\pi/16)}{3}\left[2\operatorname{sen}\frac{\pi}{16}-\operatorname{sen}\frac{\pi}{8}+2\operatorname{sen}\frac{3\pi}{16}\right] \approx 0.29285866$$

$$n=3: \quad \frac{5(\pi/20)}{24}\left[11\operatorname{sen}\frac{\pi}{20}+\operatorname{sen}\frac{\pi}{10}+\operatorname{sen}\frac{3\pi}{20}+11\operatorname{sen}\frac{\pi}{5}\right] \approx 0.29286923$$

La tabla 4.8 resume los resultados y muestra los errores de aproximación. ■

Tabla 4.8

n	0	1	2	3	4
Fórmulas cerradas		0.27768018	0.29293264	0.29291070	0.29289318
Error		0.01521303	0.00003942	0.00001748	0.00000004
Fórmulas abiertas	0.30055887	0.29798754	0.29285866	0.29286923	
Error	0.00766565	0.00509432	0.00003456	0.00002399	

La sección Conjunto de ejercicios 4.3 está disponible en línea. Encuentre la ruta de acceso en las páginas preliminares.

4.4 Integración numérica compuesta

En general, el uso de fórmulas de Newton-Cotes es inapropiado sobre largos intervalos de integración. Se requerirían fórmulas de grado superior y los valores de los coeficientes en estas fórmulas son difíciles de obtener. Además, las fórmulas de Newton-Cotes están basadas en los polinomios de interpolación que utilizan nodos igualmente espaciados, un procedimiento inapropiado sobre intervalos largos debido a la naturaleza oscilatoria de los polinomios de orden superior.

A menudo, la aproximación por tramos es efectiva. Recuerde que esto se usó para interpolación de spline.

En esta sección, analizamos el enfoque por *tramos* (o fragmentario) para la integración numérica que usa las fórmulas de Newton-Cotes de bajo orden. Estas son las técnicas que se aplican más a menudo.

Ejemplo 1 Use la regla de Simpson para aproximar $\int_0^4 e^x\,dx$ y compare esto con los resultados obtenidos mediante la suma de las aproximaciones de Simpson para $\int_0^2 e^x\,dx$ y $\int_2^4 e^x\,dx$ y al sumar éstas con $\int_0^1 e^x\,dx$, $\int_1^2 e^x\,dx$, $\int_2^3 e^x\,dx$, y $\int_3^4 e^x\,dx$.

Solución La regla de Simpson en [0, 4] con $h=2$ da

$$\int_0^4 e^x\,dx \approx \frac{2}{3}(e^0+4e^2+e^4)=56.76958.$$

La respuesta correcta en este caso es $e^4-e^0=53.59815$, y el error -3.17143 es mucho más grande de lo que aceptaríamos normalmente.

Al aplicar la regla de Simpson en cada uno de los intervalos [0, 2] y [2, 4] con $h = 1$ da

$$\begin{aligned}\int_0^4 e^x\,dx &= \int_0^2 e^x\,dx + \int_2^4 e^x\,dx \\ &\approx \frac{1}{3}\left(e^0 + 4e + e^2\right) + \frac{1}{3}\left(e^2 + 4e^3 + e^4\right) \\ &= \frac{1}{3}\left(e^0 + 4e + 2e^2 + 4e^3 + e^4\right) \\ &= 53.86385.\end{aligned}$$

El error se ha reducido a -0.26570.

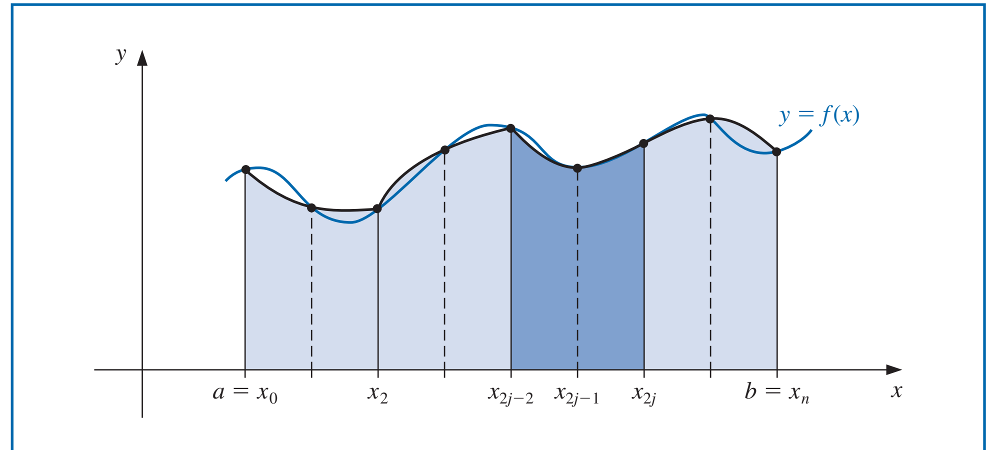
Para las integrales en [0, 1], [1, 2], [3, 4], y [3, 4], utilizamos la regla de Simpson cuatro veces con $h = \frac{1}{2}$, con lo que obtenemos

$$\begin{aligned}\int_0^4 e^x\,dx &= \int_0^1 e^x\,dx + \int_1^2 e^x\,dx + \int_2^3 e^x\,dx + \int_3^4 e^x\,dx \\ &\approx \frac{1}{6}\left(e_0 + 4e^{1/2} + e\right) + \frac{1}{6}\left(e + 4e^{3/2} + e^2\right) \\ &\quad + \frac{1}{6}\left(e^2 + 4e^{5/2} + e^3\right) + \frac{1}{6}\left(e^3 + 4e^{7/2} + e^4\right) \\ &= \frac{1}{6}\left(e^0 + 4e^{1/2} + 2e + 4e^{3/2} + 2e^2 + 4e^{5/2} + 2e^3 + 4e^{7/2} + e^4\right) \\ &= 53.61622.\end{aligned}$$

El error para esta aproximación se ha reducido a -0.01807. ■

Para generalizar este procedimiento para una integral arbitraria $\int_a^b f(x)\,dx$, seleccione un entero par n. Subdivida el intervalo $[a, b]$ en n subintervalos y aplique la regla de Simpson en cada par consecutivo de subintervalos (véase la figura 4.7).

Figura 4.7



Con $h = (b-a)/n$ y $x_j = a + jh$, para cada $j = 0, 1, \ldots, n$, tenemos

$$\begin{aligned}\int_a^b f(x)\,dx &= \sum_{j=1}^{n/2} \int_{x_{2j-2}}^{x_{2j}} f(x)\,dx \\ &= \sum_{j=1}^{n/2} \left\{ \frac{h}{3}[f(x_{2j-2}) + 4f(x_{2j-1}) + f(x_{2j})] - \frac{h^5}{90} f^{(4)}(\xi_j) \right\},\end{aligned}$$

para algunos ξ_j con $x_{2j-2} < \xi_j < x_{2j}$, siempre que $f \in C^4[a,b]$. A través del hecho de que para cada $j = 1, 2, \ldots, (n/2) - 1$ tenemos $f(x_{2j})$ que figura en el término correspondiente al intervalo $[x_{2j-2}, x_{2j}]$ y también en el término correspondiente al intervalo $[x_{2j}, x_{2j+2}]$, podemos reducir esta suma a

$$\int_a^b f(x)\,dx = \frac{h}{3}\left[f(x_0) + 2\sum_{j=1}^{(n/2)-1} f(x_{2j}) + 4\sum_{j=1}^{n/2} f(x_{2j-1}) + f(x_n)\right] - \frac{h^5}{90}\sum_{j=1}^{n/2} f^{(4)}(\xi_j).$$

El error relacionado con esta aproximación es

$$E(f) = -\frac{h^5}{90}\sum_{j=1}^{n/2} f^{(4)}(\xi_j),$$

donde $x_{2j-2} < \xi_j < x_{2j}$, para cada $j = 1, 2, \ldots, n/2$.

Si $f \in C^4[a,b]$, el teorema de valor extremo 1.9 implica que $f^{(4)}$ asume su máximo y su mínimo en $[a, b]$. Puesto que

$$\min_{x\in[a,b]} f^{(4)}(x) \leq f^{(4)}(\xi_j) \leq \max_{x\in[a,b]} f^{(4)}(x),$$

tenemos

$$\frac{n}{2}\min_{x\in[a,b]} f^{(4)}(x) \leq \sum_{j=1}^{n/2} f^{(4)}(\xi_j) \leq \frac{n}{2}\max_{x\in[a,b]} f^{(4)}(x)$$

y

$$\min_{x\in[a,b]} f^{(4)}(x) \leq \frac{2}{n}\sum_{j=1}^{n/2} f^{(4)}(\xi_j) \leq \max_{x\in[a,b]} f^{(4)}(x).$$

Por medio del teorema de valor intermedio 1.11, existe $\mu \in (a, b)$ tal que

$$f^{(4)}(\mu) = \frac{2}{n}\sum_{j=1}^{n/2} f^{(4)}(\xi_j).$$

Por lo tanto,

$$E(f) = -\frac{h^5}{90}\sum_{j=1}^{n/2} f^{(4)}(\xi_j) = -\frac{h^5}{180} n f^{(4)}(\mu),$$

Y, puesto que $h = (b-a)/n$,

$$E(f) = -\frac{(b-a)}{180} h^4 f^{(4)}(\mu).$$

Estas observaciones producen el siguiente resultado.

Teorema 4.4 Si $f \in C^4[a,b]$, n es par, $h = (b-a)/n$, y $x_j = a + jh$, para cada $j = 0, 1, \ldots, n$. Existe $\mu \in (a,b)$ para los que la **regla compuesta de Simpson** para n subintervalos se puede reescribir con su término de error como

$$\int_a^b f(x)\,dx = \frac{h}{3}\left[f(a) + 2\sum_{j=1}^{(n/2)-1} f(x_{2j}) + 4\sum_{j=1}^{n/2} f(x_{2j-1}) + f(b)\right] - \frac{b-a}{180} h^4 f^{(4)}(\mu).$$

■

Observe que el término de error para la regla compuesta de Simpson es $O(h^4)$, mientras que era $O(h^5)$ para la regla estándar de Simpson. Sin embargo, estos índices no son comparables ya que para la regla estándar de Simpson, tenemos h fija en $h = (b - a)/2$, pero para la regla compuesta de Simpson, tenemos $h = (b-a)/n$, para n un entero par. Esto nos permite reducir considerablemente el valor de h.

El algoritmo 4.1 utiliza la regla compuesta de Simpson en n subintervalos. Éste es el algoritmo de cuadratura que se usa con mayor frecuencia para propósito general.

ALGORITMO **4.1**

Regla compuesta de Simpson

Para aproximar la integral $I = \int_a^b f(x)\,dx$:

ENTRADA extremos a, b; entero positivo n par.

SALIDA aproximación XI para I.

Paso 1 Haga $h = (b - a)/n$.

Paso 2 Haga $XI0 = f(a) + f(b)$;
$XI1 = 0$; (*Suma de* $f(x_{2i-1})$.)
$XI2 = 0$. (*Suma de* $f(x_{2i})$.)

Paso 3 Para $i = 1, \ldots, n-1$ realice los pasos 4 y 5.

Paso 4 Haga $X = a + ih$.

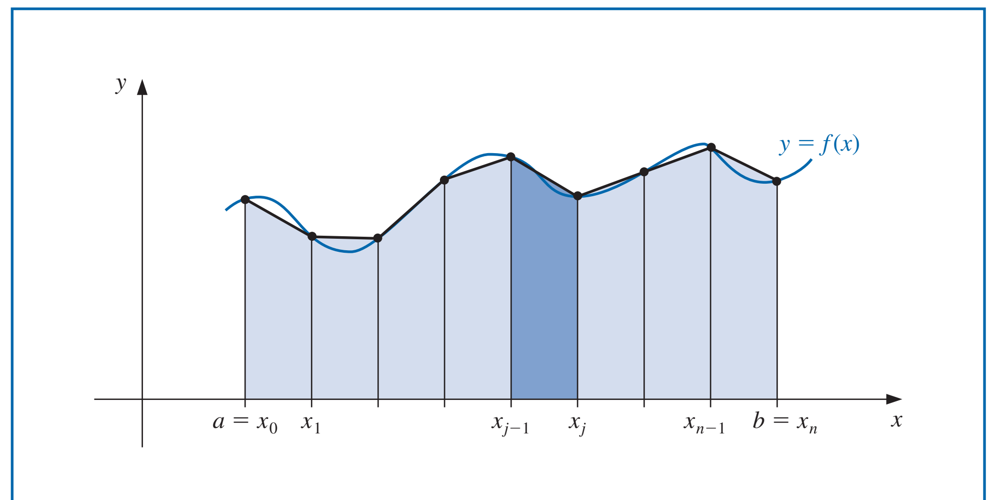
Paso 5 Si i es par entonces haga $XI2 = XI2 + f(X)$
También determine $XI1 = XI1 + f(X)$.

Paso 6 Haga $XI = h(XI0 + 2 \cdot XI2 + 4 \cdot XI1)/3$.

Paso 7 SALIDA (XI);
PARE. ■

El enfoque de subdivisión se puede aplicar a cualquiera de las fórmulas de Newton-Cotes. Las extensiones de las reglas trapezoidal (véase la figura 4.8) y de punto medio se dan sin prueba. La regla trapezoidal sólo requiere un intervalo para cada aplicación, por lo que el entero n puede ser tanto par como impar.

Figura 4.8

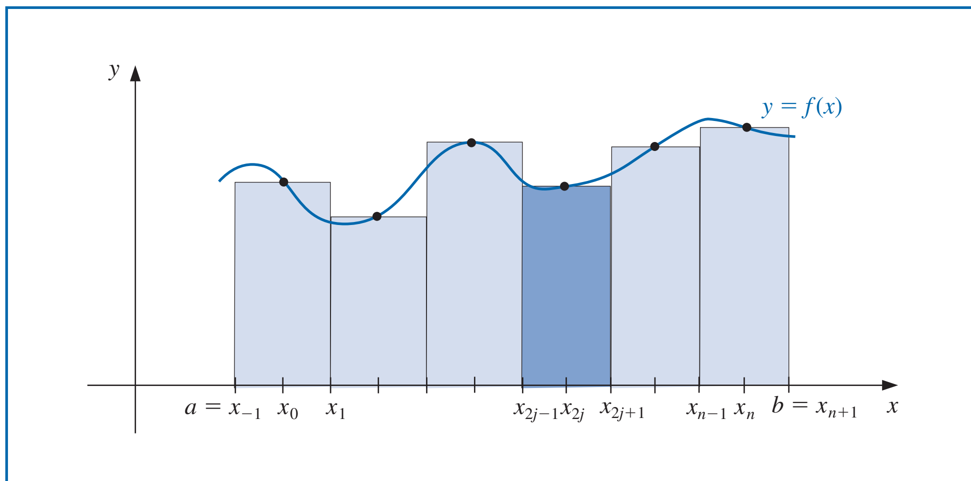


Teorema 4.5 Sean $f \in C^2[a, b]$, $h = (b - a)/n$, y $x_j = a + jh$, para cada $j = 0, 1, \ldots, n$. Existe $\mu \in (a, b)$ para el que la **regla compuesta trapezoidal** para n subintervalos se puede reescribir con este término de error como

$$\int_a^b f(x)\,dx = \frac{h}{2}\left[f(a) + 2\sum_{j=1}^{n-1} f(x_j) + f(b)\right] - \frac{b-a}{12}h^2 f''(\mu). \qquad \blacksquare$$

Para la **regla compuesta de punto medio**, n debe ser, nuevamente, par (véase la figura 4.9.)

Figura 4.9



Teorema 4.6 Si $f \in C^2[a, b]$, n es par, $h = (b - a)/(n + 2)$, y $x_j = a + (j + 1)h$ para cada $j = -1, 0, \ldots, n + 1$. Existe $\mu \in (a, b)$ para el que la **regla compuesta de punto medio** para $n + 2$ subintervalos se puede reescribir con su término de error como

$$\int_a^b f(x)\,dx = 2h\sum_{j=0}^{n/2} f(x_{2j}) + \frac{b-a}{6}h^2 f''(\mu). \qquad \blacksquare$$

Ejemplo 2 Determine valores de h que garantizarán un error de aproximación menor que 0.00002 al aproximar $\int_0^{\pi}$ sen $x\,dx$ y usar
a) La regla compuesta trapezoidal y **b)** la regla compuesta de Simpson.

Solución **a)** La forma de error para la regla trapezoidal compuesta para $f(x) = \text{sen}\,x$ en $[0, \pi]$ es

$$\left|\frac{\pi h^2}{12} f''(\mu)\right| = \left|\frac{\pi h^2}{12}(-\text{sen}\,\mu)\right| = \frac{\pi h^2}{12}|\text{sen}\,\mu|.$$

Para garantizar suficiente precisión con esta técnica, necesitamos tener

$$\frac{\pi h^2}{12}|\text{sen}\,\mu| \le \frac{\pi h^2}{12} < 0.00002.$$

Puesto que $h = \pi/n$, necesitamos

$$\frac{\pi^3}{12n^2} < 0.00002, \quad \text{lo cual implica que} \quad n > \left(\frac{\pi^3}{12(0.00002)}\right)^{1/2} \approx 359.44,$$

y la regla compuesta trapezoidal requiere $n \geq 360$.

b) La forma de error para la regla compuesta de Simpson para $f(x) = \text{sen}\, x$ en $[0, \pi]$ es

$$\left|\frac{\pi h^4}{180} f^{(4)}(\mu)\right| = \left|\frac{\pi h^4}{180} \text{sen}\, \mu\right| = \frac{\pi h^4}{180}|\text{sen}\, \mu|.$$

Para garantizar suficiente precisión con esta técnica, necesitamos tener

$$\frac{\pi h^4}{180}|\text{sen}\, \mu| \leq \frac{\pi h^4}{180} < 0.00002.$$

A través del hecho de que $n = \pi/h$ nos da

$$\frac{\pi^5}{180n^4} < 0.00002, \quad \text{lo cual implica que} \quad n > \left(\frac{\pi^5}{180(0.00002)}\right)^{1/4} \approx 17.07.$$

Por lo tanto, la regla compuesta de Simpson sólo requiere $n \geq 18$.

La regla compuesta de Simpson con $n = 18$ nos da

$$\int_0^{\pi} \text{sen}\, x\, dx \approx \frac{\pi}{54}\left[2\sum_{j=1}^{8} \text{sen}\left(\frac{j\pi}{9}\right) + 4\sum_{j=1}^{9} \text{sen}\left(\frac{(2j-1)\pi}{18}\right)\right] = 2.0000104.$$

Esto es preciso dentro de aproximadamente 10^{-5} porque el valor verdadero es $-\cos(\pi) - (-\cos(0)) = 2$. ■

La regla compuesta de Simpson es la selección clara si desea minimizar los cálculos. Para propósitos de comparación, considere la regla compuesta trapezoidal por medio de $h = \pi/18$ para la integral en el ejemplo 2. Esta aproximación utiliza las mismas evaluaciones de función que la regla compuesta de Simpson, pero la aproximación en este caso,

$$\int_0^{\pi} \text{sen}\, x\, dx \approx \frac{\pi}{36}\left[2\sum_{j=1}^{17} \text{sen}\left(\frac{j\pi}{18}\right) + \text{sen}\, 0 + \text{sen}\, \pi\right] = \frac{\pi}{36}\left[2\sum_{j=1}^{17} \text{sen}\left(\frac{j\pi}{18}\right)\right]$$
$$= 1.9949205,$$

sólo es precisa para aproximadamente 5×10^{-3}.

Estabilidad del error de redondeo

En el ejemplo 2, observamos que garantizar una precisión de 2×10^{-5} para aproximar $\int_0^{\pi} \text{sen}\, x\, dx$ requería 360 subdivisiones de $[0, \pi]$ para la regla compuesta trapezoidal y sólo 18 para la regla compuesta de Simpson. Además del hecho de que se necesitan menos cálculos para la técnica de Simpson, usted podría sospechar que este método también implica menos error de redondeo. Sin embargo, una propiedad importante compartida por todas las técnicas de integración compuesta es una estabilidad respecto al error de redondeo. Es decir, el error de redondeo no depende del número cálculos realizados.

Se espera que la integración numérica sea estable, mientras que la diferenciación numérica es inestable.

Para demostrar este hecho considerablemente sorprendente, suponga que aplicamos la regla compuesta de Simpson con n subintervalos a una función f en $[a, b]$ y determine la máxima cota para el error de redondeo. Suponga que $f(x_i)$ se aproxima mediante $\tilde{f}(x_i)$ y que

$$f(x_i) = \tilde{f}(x_i) + e_i, \quad \text{para cada } i = 0, 1, \ldots, n,$$

donde e_i denota el error de redondeo asociado con el uso de $\tilde{f}(x_i)$ para aproximar $f(x_i)$. Entonces, el error acumulado, $e(h)$, en la regla compuesta de Simpson es

$$\begin{aligned} e(h) &= \left| \frac{h}{3}\left[e_0 + 2\sum_{j=1}^{(n/2)-1} e_{2j} + 4\sum_{j=1}^{n/2} e_{2j-1} + e_n \right] \right| \\ &\le \frac{h}{3}\left[|e_0| + 2\sum_{j=1}^{(n/2)-1} |e_{2j}| + 4\sum_{j=1}^{n/2} |e_{2j-1}| + |e_n| \right]. \end{aligned}$$

Si los errores de redondeo están limitados de manera uniforme por ε, entonces

$$e(h) \le \frac{h}{3}\left[\varepsilon + 2\left(\frac{n}{2} - 1\right)\varepsilon + 4\left(\frac{n}{2}\right)\varepsilon + \varepsilon \right] = \frac{h}{3} 3n\varepsilon = nh\varepsilon.$$

Sin embargo, $nh = b - a$, por lo que

$$e(h) \le (b-a)\varepsilon,$$

una cota independiente de h (y n). Esto significa que, aunque quizá necesitemos dividir un intervalo en más partes para garantizar precisión, el cálculo incrementado que se requiere no aumenta el error de redondeo. Este resultado implica que el procedimiento es estable conforme h se aproxima a cero. Recuerde que esto no es verdad para los procedimientos de diferenciación numérica considerados al principio de este capítulo.

La sección Conjunto de ejercicios 4.4 está disponible en línea. Encuentre la ruta de acceso en las páginas preliminares.

4.5 Integración de Romberg

En esta sección ilustraremos la forma en la que la extrapolación de Richardson aplicada a los resultados de la regla compuesta trapezoidal se puede usar para obtener aproximaciones de alta precisión con poco costo computacional.

En la sección 4.4 encontramos que la regla compuesta trapezoidal tiene un error de truncamiento de orden $O(h^2)$. En específico, mostramos que para $h = (b-a)/n$ y $x_j = a + jh$, tenemos

$$\int_a^b f(x)\,dx = \frac{h}{2}\left[f(a) + 2\sum_{j=1}^{n-1} f(x_j) + f(b) \right] - \frac{(b-a)f''(\mu)}{12}h^2,$$

para algún numero μ en (a, b).

Con un método alternativo, se puede mostrar (consulte [RR], pp. 136–140) que si $f \in C^\infty[a, b]$, la regla compuesta trapezoidal también se puede escribir con un término de error de la forma

$$\int_a^b f(x)\,dx = \frac{h}{2}\left[f(a) + 2\sum_{j=1}^{n-1} f(x_j) + f(b) \right] + K_1h^2 + K_2h^4 + K_3h^6 + \cdots, \tag{4.33}$$

donde K_i es una constante que depende solamente de $f^{(2i-1)}(a)$ y $f^{(2i-1)}(b)$.

Recuerde, de la sección 4.2, que la extrapolación de Richardson se puede realizar en cualquier procedimiento de aproximación cuyo error de truncamiento es de la forma

$$\sum_{j=1}^{m-1} K_j h^{\alpha_j} + O(h^{\alpha_m}),$$

para un conjunto de constantes K_j y donde $\alpha_1 < \alpha_2 < \alpha_3 < \cdots < \alpha_m$. En esa sección realizamos demostraciones para ilustrar qué tan efectiva es esta técnica cuando el procedimiento de aproximación tiene un error de truncamiento sólo con potencias pares de h, es decir, cuando el error de truncamiento tiene la forma

$$\sum_{j=1}^{m-1} K_j h^{2j} + O(h^{2m}).$$

Werner Romberg (1909–2093) concibió este procedimiento para mejorar la precisión de la regla trapezoidal al eliminar términos sucesivos en la expansión asintótica en 1955.

Puesto que la regla trapezoidal compuesta tiene esta forma, es un candidato obvio para extrapolación. Esto resulta en una técnica conocida como **integración de Romberg**.

Para aproximar la integral $\int_a^b f(x)\,dx$, utilizamos los resultados de la regla compuesta trapezoidal con $n = 1, 2, 4, 8, 16, \ldots,$ e indicamos las aproximaciones resultantes, respectivamente mediante $R_{1,1}$, $R_{2,1}$, $R_{3,1}$, y así sucesivamente. A continuación, aplicamos la extrapolación de la forma determinada en la sección 4.2; es decir, obtenemos $O(h^4)$ aproximaciones $R_{2,2}$, $R_{3,2}$, $R_{4,2}$, y así sucesivamente, por medio de

$$R_{k,2} = R_{k,1} + \frac{1}{3}(R_{k,1} - R_{k-1,1}), \quad \text{para } k = 2, 3, \ldots$$

y, entonces, las $O(h^6)$ aproximaciones $R_{3,3}$, $R_{4,3}$, $R_{5,3}$, se pueden obtener mediante

$$R_{k,3} = R_{k,2} + \frac{1}{15}(R_{k,2} - R_{k-1,2}), \quad \text{para } k = 3, 4, \ldots$$

En general, después de que se han obtenido las aproximaciones $R_{k,j-1}$ adecuadas, determinamos las aproximaciones $O(h^{2j})$ a partir de

$$R_{k,j} = R_{k,j-1} + \frac{1}{4^{j-1}-1}(R_{k,j-1} - R_{k-1,j-1}), \quad \text{para } k = j, j+1, \ldots$$

Ejemplo 1 Use la regla compuesta trapezoidal para encontrar aproximaciones para $\int_0^\pi \operatorname{sen} x\,dx$ con $n = 1, 2, 4, 8$ y 16. A continuación, realice la integración de Romberg en los resultados.

La regla compuesta trapezoidal para los diferentes valores de n proporciona las siguientes aproximaciones para el valor verdadero 2:

$$R_{1,1} = \frac{\pi}{2}[\operatorname{sen} 0 + \operatorname{sen} \pi] = 0,$$

$$R_{2,1} = \frac{\pi}{4}\left[\operatorname{sen} 0 + 2\operatorname{sen}\frac{\pi}{2} + \operatorname{sen}\pi\right] = 1.57079633,$$

$$R_{3,1} = \frac{\pi}{8}\left[\operatorname{sen} 0 + 2\left(\operatorname{sen}\frac{\pi}{4} + \operatorname{sen}\frac{\pi}{2} + \operatorname{sen}\frac{3\pi}{4}\right) + \operatorname{sen}\pi\right] = 1.89611890,$$

$$R_{4,1} = \frac{\pi}{16}\left[\operatorname{sen} 0 + 2\left(\operatorname{sen}\frac{\pi}{8} + \operatorname{sen}\frac{\pi}{4} + \cdots + \operatorname{sen}\frac{3\pi}{4} + \operatorname{sen}\frac{7\pi}{8}\right) + \operatorname{sen}\pi\right]$$

$$= 1.97423160, \text{ y}$$

$$R_{5,1} = \frac{\pi}{32}\left[\operatorname{sen} 0 + 2\left(\operatorname{sen}\frac{\pi}{16} + \operatorname{sen}\frac{\pi}{8} + \cdots + \operatorname{sen}\frac{7\pi}{8} + \operatorname{sen}\frac{15\pi}{16}\right) + \operatorname{sen}\pi\right]$$

$$= 1.99357034.$$

Las aproximaciones $O(h^4)$ son

$$R_{2,2} = R_{2,1} + \frac{1}{3}(R_{2,1} - R_{1,1}) = 2.09439511,$$

$$R_{3,2} = R_{3,1} + \frac{1}{3}(R_{3,1} - R_{2,1}) = 2.00455976,$$

$$R_{4,2} = R_{4,1} + \frac{1}{3}(R_{4,1} - R_{3,1}) = 2.00026917, \text{ y}$$

$$R_{5,2} = R_{5,1} + \frac{1}{3}(R_{5,1} - R_{4,1}) = 2.00001659.$$

Las aproximaciones $O(h^6)$ son

$$R_{3,3} = R_{3,2} + \frac{1}{15}(R_{3,2} - R_{2,2}) = 1.99857073,$$

$$R_{4,3} = R_{4,2} + \frac{1}{15}(R_{4,2} - R_{3,2}) = 1.99998313, \text{ y}$$

$$R_{5,3} = R_{5,2} + \frac{1}{15}(R_{5,2} - R_{4,2}) = 1.99999975.$$

Las dos aproximaciones $O(h^8)$ son

$$R_{4,4} = R_{4,3} + \frac{1}{63}(R_{4,3} - R_{3,3}) = 2.00000555 \text{ y } R_{5,4} = R_{5,3} + \frac{1}{63}(R_{5,3} - R_{4,3}) = 2.00000001,$$

y la aproximación $O(h^{10})$ final es

$$R_{5,5} = R_{5,4} + \frac{1}{255}(R_{5,4} - R_{4,4}) = 1.99999999.$$

Estos resultados se muestran en la tabla 4.9. ■

Tabla 4.9

0				
1.57079633	2.09439511			
1.89611890	2.00455976	1.99857073		
1.97423160	2.00026917	1.99998313	2.00000555	
1.99357034	2.00001659	1.99999975	2.00000001	1.99999999

Observe que al generar las aproximaciones para la regla compuesta trapezoidal en el ejemplo 1, cada aproximación consecutiva incluía todas las evaluaciones de la función desde la aproximación previa. Es decir, $R_{1,1}$ utilizaba evaluaciones en 0 y π, y $R_{2,1}$ usaba estas evaluaciones y añadía una evaluación en el punto intermedio $\pi/2$. Entonces $R_{3,1}$ utilizaba las evaluaciones de $R_{2,1}$ y añadía dos intermedios adicionales en $\pi/4$ y $3\pi/4$. Este patrón continúa con $R_{4,1}$ mediante las mismas evaluaciones como $R_{3,1}$, pero al añadir evaluaciones en los cuatro puntos intermedios $\pi/8, 3\pi/8, 5\pi/8, 7\pi/8,$ y así sucesivamente.

Este procedimiento de evaluación para las aproximaciones de la regla compuesta tradicional se mantiene para una integral en cualquier intervalo $[a, b]$. En general, la regla compuesta trapezoidal denotaba $R_{k+1,1}$ utiliza las mismas evaluaciones que $R_{k,1}$, pero añade evaluaciones en los puntos intermedios 2^{k-2}. Por lo tanto, el cálculo eficiente de estas aproximaciones puede realizarse de manera recursiva.

Para obtener las aproximaciones de la regla compuesta trapezoidal para $\int_a^b f(x)\,dx$, entonces $h_k = (b-a)/m_k = (b-a)/2^{k-1}$.

$$R_{1,1} = \frac{h_1}{2}[f(a) + f(b)] = \frac{(b-a)}{2}[f(a) + f(b)],$$

y

$$R_{2,1} = \frac{h_2}{2}[f(a) + f(b) + 2f(a + h_2)].$$

Al reescribir este resultado para $R_{2,1}$, podemos incorporar la aproximación previamente determinada $R_{1,1}$

$$R_{2,1} = \frac{(b-a)}{4}\left[f(a) + f(b) + 2f\left(a + \frac{(b-a)}{2}\right)\right] = \frac{1}{2}[R_{1,1} + h_1 f(a + h_2)].$$

De manera similar, podemos escribir

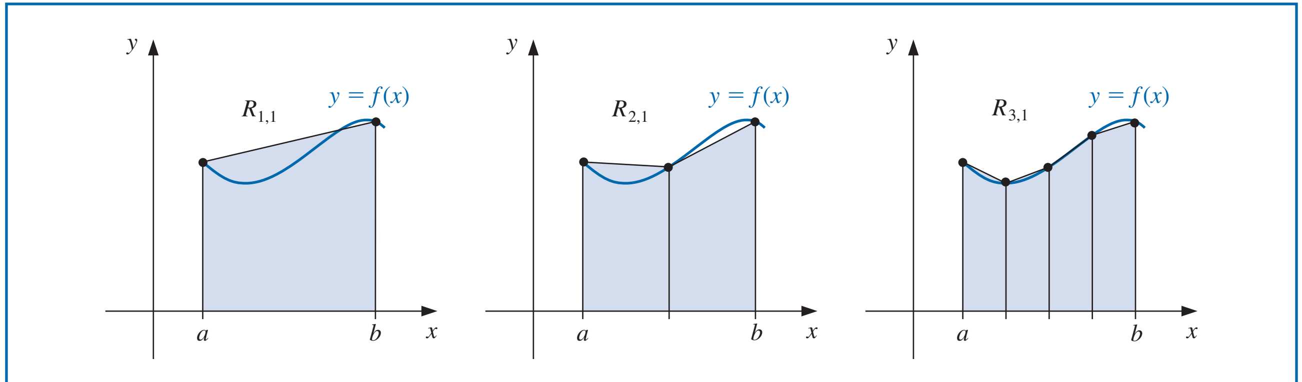
$$R_{3,1} = \frac{1}{2}\{R_{2,1} + h_2[f(a + h_3) + f(a + 3h_3)]\},$$

y, en general (véase la figura 4.10), tenemos

$$R_{k,1} = \frac{1}{2}\left[R_{k-1,1} + h_{k-1}\sum_{i=1}^{2^{k-2}} f(a + (2i-1)h_k)\right], \tag{4.34}$$

para cada $k = 2, 3, \ldots, n$. (Consulte los ejercicios 18 y 19.)

Figura 4.10



Entonces, la extrapolación se usa para producir aproximaciones $O(h_k^{2j})$ mediante

$$R_{k,j} = R_{k,j-1} + \frac{1}{4^{j-1}-1}(R_{k,j-1} - R_{k-1,j-1}), \quad \text{para } k = j, j+1, \ldots,$$

como se muestra en la tabla 4.10.

Tabla 4.10

k	$O(h_k^2)$	$O(h_k^4)$	$O(h_k^6)$	$O(h_k^8)$		$O(h_k^{2n})$
1	$R_{1,1}$					
2	$R_{2,1}$	$R_{2,2}$				
3	$R_{3,1}$	$R_{3,2}$	$R_{3,3}$			
4	$R_{4,1}$	$R_{4,2}$	$R_{4,3}$	$R_{4,4}$		
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	
n	$R_{n,1}$	$R_{n,2}$	$R_{n,3}$	$R_{n,4}$	$\cdots$	$R_{n,n}$

El método efectivo para construir la tabla de Romberg utiliza el orden más alto de la aproximación en cada paso. Es decir, calcula las entradas fila por fila, en orden $R_{1,1}$, $R_{2,1}$, $R_{2,2}$, $R_{3,1}$, $R_{3,2}$, $R_{3,3}$, y así sucesivamente. Esto también permite calcular una nueva fila completa en la tabla al realizar sólo una aplicación adicional de la regla compuesta trapezoidal. A continuación, usa un promediado simple de los valores previamente calculados para obtener las entradas restantes en la fila. Recuerde

- Calcular la tabla de Romberg una fila completa a la vez.

Ejemplo 2 Añada una fila de extrapolación adicional a la tabla 4.9 para $\int_0^{\pi} \operatorname{sen} x\, dx$.

Solución Para obtener la fila adicional, necesitamos la aproximación trapezoidal

$$R_{6,1} = \frac{1}{2}\left[R_{5,1} + \frac{\pi}{16}\sum_{k=1}^{2^4} \operatorname{sen} \frac{(2k-1)\pi}{32}\right] = 1.99839336.$$

El valor en la tabla 4.9 da

$$\begin{aligned}
R_{6,2} &= R_{6,1} + \frac{1}{3}(R_{6,1} - R_{5,1}) = 1.99839336 + \frac{1}{3}(1.99839336 - 1.99357035)\\
&= 2.00000103,\\
R_{6,3} &= R_{6,2} + \frac{1}{15}(R_{6,2} - R_{5,2}) = 2.00000103 + \frac{1}{15}(2.00000103 - 2.00001659)\\
&= 2.00000000,\\
R_{6,4} &= R_{6,3} + \frac{1}{63}(R_{6,3} - R_{5,3}) = 2.00000000, \quad R_{6,5} = R_{6,4} + \frac{1}{255}(R_{6,4} - R_{5,4})\\
&= 2.00000000,
\end{aligned}$$

y $R_{6,6} = R_{6,5} + \frac{1}{1023}(R_{6,5} - R_{5,5}) = 2.00000000$. La nueva tabla de extrapolación se muestra en la tabla 4.11. ■

Tabla 4.11

0					
1.57079633	2.09439511				
1.89611890	2.00455976	1.99857073			
1.97423160	2.00026917	1.99998313	2.00000555		
1.99357034	2.00001659	1.99999975	2.00000001	1.99999999	
1.99839336	2.00000103	2.00000000	2.00000000	2.00000000	2.00000000

Observe que todos los valores extrapolados excepto por el primero (en la primera fila de la segunda columna) son más precisos que la mejor aproximación compuesta trapezoidal (en la última fila de la primera columna). A pesar de que existen 21 entradas en la tabla 4.11, sólo la sexta en la columna izquierda requiere evaluaciones de función ya que son las únicas entradas generadas por la regla compuesta trapezoidal; las otras entradas se obtienen mediante un proceso de promediado. De hecho, debido a la relación de recurrencia de los términos en la columna izquierda, las únicas evaluaciones de función son aquellas para calcular la aproximación de la regla compuesta trapezoidal. En general, $R_{k,1}$ requiere evaluaciones de función $1 + 2^{k-1}$, por lo que en este caso se necesita $1 + 2^5 = 33$.

El algoritmo 4.2 usa el procedimiento recursivo para encontrar las aproximaciones de la regla compuesta trapezoidal y calcula los resultados en la tabla fila por fila.

ALGORITMO 4.2

Integración de Romberg

Para aproximar la integral $I = \int_a^b f(x)\,dx$, seleccione un entero $n > 0$.

ENTRADA extremos a, b; entero n.

SALIDA un arreglo R. (*Calcule R por filas; sólo se almacenan las últimas dos filas*.)

Paso 1 Haga $h = b - a$;
$R_{1,1} = \frac{h}{2}(f(a) + f(b))$.

Paso 2 SALIDA $(R_{1,1})$.

Paso 3 Para $i = 2, \ldots, n$ realice los pasos 4–8.

Paso 4 Haga $R_{2,1} = \frac{1}{2}\left[R_{1,1} + h\sum_{k=1}^{2^{i-2}} f(a + (k - 0.5)h)\right]$.

(*Aproximación a partir del método trapezoidal*.)

Paso 5 Para $j = 2, \ldots, i$
haga $R_{2,j} = R_{2,j-1} + \frac{R_{2,j-1} - R_{1,j-1}}{4^{j-1} - 1}$. (*Extrapolación*.)

Paso 6 SALIDA ($R_{2,j}$ para $j = 1, 2, \ldots, i$).

Paso 7 Haga $h = h/2$.

Paso 8 Para $j = 1, 2, \ldots, i$ determine $R_{1,j} = R_{2,j}$. (*Actualice la fila 1 de R*.)

Paso 9 PARE. ■

El algoritmo 4.2 requiere un entero preestablecido n para determinar el número de filas que se va a generar. También estableceremos una tolerancia de error para la aproximación y generar n, dentro de algúna cota superior hasta que las entradas diagonales consecutivas $R_{n-1,n-1}$ y $R_{n,n}$ concuerden dentro de la tolerancia. Para evitar la posibilidad de que dos elementos de fila consecutivos concuerden unos con otros, pero no con el valor de la integral que se está aproximando, es común generar aproximaciones hasta que no sólo $|R_{n-1,n-1} - R_{n,n}|$ esté dentro de la tolerancia, sino también $|R_{n-2,n-2} - R_{n-1,n-1}|$. A pesar de que no es una protección universal, esto garantizará que dos conjuntos de aproximaciones generados de manera diferente concuerden dentro de la tolerancia especificada antes de $R_{n,n}$ se acepte como suficientemente preciso.

La integración de Romberg aplicada a una función f en el intervalo $[a, b]$ depende de la suposición de que la regla compuesta trapezoidal tienen un término de error que se puede expresar en la forma de la ecuación (4.33); es decir, debemos tener $f \in C^{2k+2}[a, b]$ para la k-ésima fila que se va a generar. Los algoritmos de propósito general por medio de la integración de Romberg incluyen una verificación en cada etapa para garantizar el cumplimiento de esta suposición. Estos métodos se conocen como *algoritmos cautelosos de Romberg* y se describen en [Joh]. Esta referencia también describe métodos para utilizar la técnica de Romberg como un procedimiento adaptable, similar a la regla adaptable de Simpson, la cual se analizará en la sección 4.6.

El adjetivo *cauteloso* que se usa en la descripción de un método numérico indica que se incluye una verificación para determinar si las hipótesis de continuidad tienen alguna probabilidad de ser verdaderas.

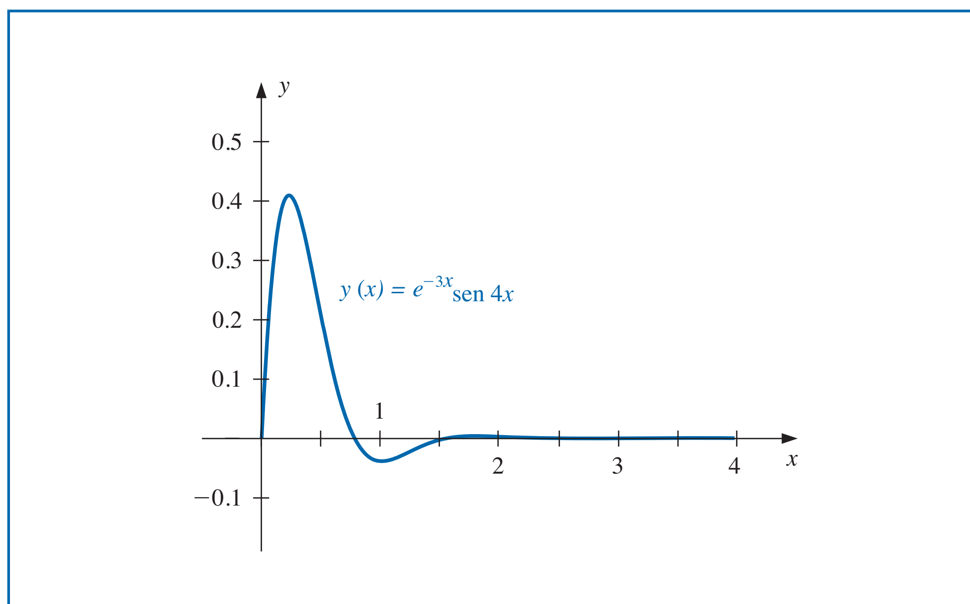
La sección Conjunto de ejercicios 4.5 está disponible en línea. Encuentre la ruta de acceso en páginas preliminares.

4.6 Métodos de cuadratura adaptable

Las fórmulas compuestas son muy efectivas en muchas situaciones, pero ocasionalmente sufren porque requieren el uso de nodos igualmente espaciados. Esto es inadecuado al integrar una función sobre un intervalo que contiene ambas regiones con gran variación funcional y regiones con variación funcional pequeña.

Ilustración La única solución a la ecuación diferencial $y'' + 6y' + 25 = 0$ que adicionalmente satisface $y(0) = 0$ y $y'(0) = 4$ es $y(x) = e^{-3x} \operatorname{sen} 4x$. Las funciones de este tipo son comunes en ingeniería mecánica porque describen ciertas características de sistemas absorbentes de muelle y amortiguador y en ingeniería eléctrica porque son soluciones comunes a los problemas fundamentales de circuitos. La gráfica de $y(x)$ para x en el intervalo [0, 4] se muestra en la figura 4.11.

Figura 4.11



Suponga que necesitamos la integral de $y(x)$ en [0, 4]. La gráfica indica que la integral en [3, 4] debe estar muy cerca de 0 y en [2, 3] tampoco se esperaría que fuera grande. Sin embargo, en [0, 2], existe variación significativa de la función y no es del todo clara cuál es la integral en este intervalo. Este es un ejemplo de una situación en donde la integración compuesta sería inadecuada. Se podría usar un método de orden bajo en [2, 4], pero se necesitaría un método de orden superior en [0, 2]. ■

La pregunta que consideraríamos en esta sección es:

- ¿Cómo podemos determinar la técnica que deberíamos aplicar en las diferentes partes del intervalo de integración y qué tan precisa podemos esperar que sea la aproximación final?

Veremos que en ciertas condiciones razonables, podemos responder esta pregunta y también determinar aproximaciones que satisfacen requisitos de precisión determinados.

Si el error de aproximación para una integral en un intervalo determinado está distribuido de manera equitativa, se necesita un tamaño de paso más pequeño para las grandes regiones de variación que para aquellas con menos variación. Una técnica eficiente para este

tipo de problema debería predecir la cantidad de variación funcional y adaptar el tamaño de paso conforme sea necesario. Estos métodos reciben el nombre de **métodos de cuadratura adaptable**. Los métodos adaptables son especialmente populares para la inclusión en paquetes profesionales de software porque, además de ser eficientes, en general proporcionan aproximaciones que se encuentran dentro de una tolerancia específica determinada.

En esta sección, consideramos un método de cuadratura y veremos cómo se puede usar para reducir el error de aproximación y también predecir un error calculado para la aproximación que no depende del conocimiento de las derivadas superiores de la función. El método que analizamos está basado en la regla compuesta de Simpson, pero la técnica se modifica fácilmente para utilizar otros procedimientos compuestos.

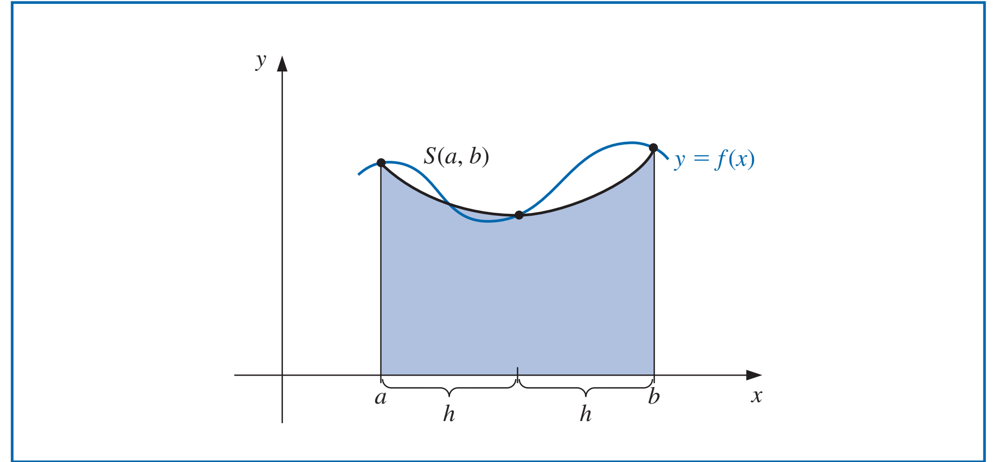
Suponga que queremos aproximar $\int_a^b f(x)\,dx$ dentro de una tolerancia específica $\varepsilon > 0$. El primer paso es aplicar la regla de Simpson con tamaño de paso $h = (b-a)/2$. Esto produce (véase la figura 4.12)

$$\int_a^b f(x)\,dx = S(a,b) - \frac{h^5}{90} f^{(4)}(\xi), \quad \text{para algunos } \xi \text{ en } (a,b), \tag{4.35}$$

donde denotamos la aproximación de la regla de Simpson en $[a, b]$ mediante

$$S(a,b) = \frac{h}{3}[f(a) + 4f(a+h) + f(b)].$$

Figura 4.12



El siguiente paso es determinar una aproximación de precisión que no requiere $f^{(4)}(\xi)$. Para hacerlo aplicamos la regla compuesta de Simpson con $n = 4$ y el tamaño de paso $(b-a)/4 = h/2$, lo cual nos da

$$\int_a^b f(x)\,dx = \frac{h}{6}\left[f(a) + 4f\left(a+\frac{h}{2}\right) + 2f(a+h) + 4f\left(a+\frac{3h}{2}\right) + f(b)\right]$$
$$- \left(\frac{h}{2}\right)^4 \frac{(b-a)}{180} f^{(4)}(\tilde{\xi}), \tag{4.36}$$

para algunos $\tilde{\xi}$ en (a, b). Para simplificar la notación, si

$$S\left(a, \frac{a+b}{2}\right) = \frac{h}{6}\left[f(a) + 4f\left(a+\frac{h}{2}\right) + f(a+h)\right]$$

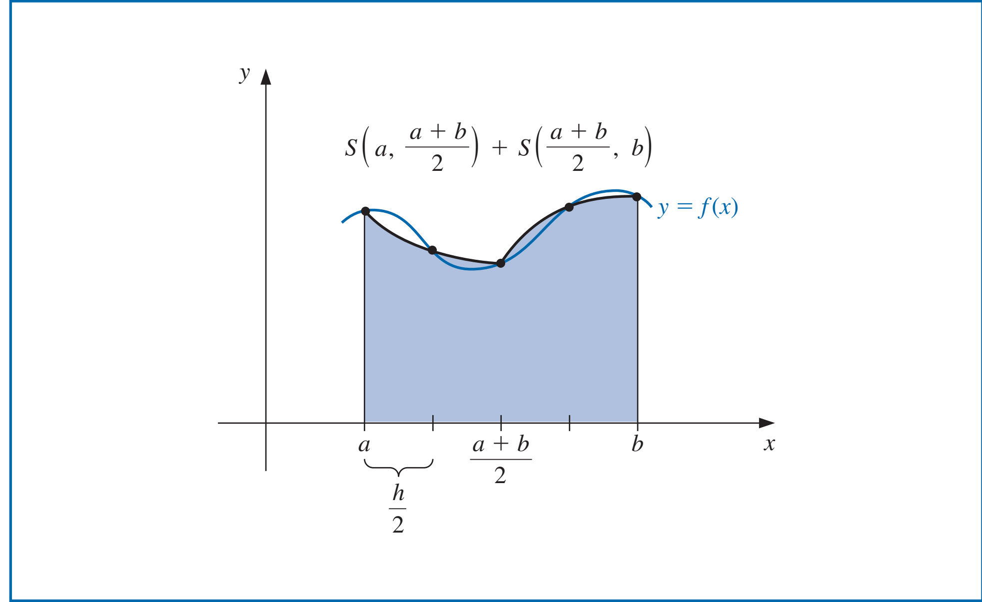
y

$$S\left(\frac{a+b}{2}, b\right) = \frac{h}{6}\left[f(a+h) + 4f\left(a + \frac{3h}{2}\right) + f(b)\right].$$

Entonces, la ecuación (4.36) se puede reescribir (véase la figura 4.13) como

$$\int_a^b f(x)\,dx = S\left(a, \frac{a+b}{2}\right) + S\left(\frac{a+b}{2}, b\right) - \frac{1}{16}\left(\frac{h^5}{90}\right) f^{(4)}(\tilde{\xi}). \qquad (4.37)$$

Figura 4.13



El cálculo de error se deriva al suponer que $\xi \approx \tilde{\xi}$ o, con mayor precisión, que $f^{(4)}(\xi) \approx f^{(4)}(\tilde{\xi})$, y el éxito de la técnica depende de la precisión de esta suposición. Si es precisa, entonces, al equiparar las integrales en las ecuaciones (4.35) y (4.37) obtenemos

$$S\left(a, \frac{a+b}{2}\right) + S\left(\frac{a+b}{2}, b\right) - \frac{1}{16}\left(\frac{h^5}{90}\right) f^{(4)}(\xi) \approx S(a,b) - \frac{h^5}{90} f^{(4)}(\xi),$$

por lo que

$$\frac{h^5}{90} f^{(4)}(\xi) \approx \frac{16}{15}\left[S(a,b) - S\left(a, \frac{a+b}{2}\right) - S\left(\frac{a+b}{2}, b\right)\right].$$

Al utilizar esta técnica en la ecuación (4.37) produce el cálculo de error

$$\left|\int_a^b f(x)\,dx - S\left(a, \frac{a+b}{2}\right) - S\left(\frac{a+b}{2}, b\right)\right|$$

$$\approx \frac{1}{16}\left(\frac{h^5}{90}\right) f^{(4)}(\xi) \approx \frac{1}{15}\left|S(a,b) - S\left(a, \frac{a+b}{2}\right) - S\left(\frac{a+b}{2}, b\right)\right|.$$

Esto implica que $S(a, (a+b)/2) + S((a+b)/2, b)$ se aproxima a $\int_a^b f(x)\,dx$ aproximadamente 15 veces mejor que lo que concuerda con el valor calculado $S(a, b)$. Por lo tanto, si

$$\left| S(a,b) - S\left(a, \frac{a+b}{2}\right) - S\left(\frac{a+b}{2}, b\right) \right| < 15\varepsilon, \tag{4.38}$$

esperamos tener

$$\left| \int_a^b f(x)\,dx - S\left(a, \frac{a+b}{2}\right) - S\left(\frac{a+b}{2}, b\right) \right| < \varepsilon, \tag{4.39}$$

y

$$S\left(a, \frac{a+b}{2}\right) + S\left(\frac{a+b}{2}, b\right)$$

se supone que es una aproximación suficientemente precisa para $\int_a^b f(x)\,dx$.

Ejemplo 1 Verifique la precisión del cálculo de error determinado por las desigualdades (4.38) y (4.39) al aplicar la integral

$$\int_0^{\pi/2} \operatorname{sen} x\,dx = 1$$

al comparar

$$\frac{1}{15}\left| S\left(0, \frac{\pi}{2}\right) - S\left(0, \frac{\pi}{4}\right) - S\left(\frac{\pi}{4}, \frac{\pi}{2}\right) \right| \text{ con } \left| \int_0^{\pi/2} \operatorname{sen} x\,dx - S\left(0, \frac{\pi}{4}\right) - S\left(\frac{\pi}{4}, \frac{\pi}{2}\right) \right|.$$

Solución Tenemos

$$S\left(0, \frac{\pi}{2}\right) = \frac{\pi/4}{3}\left[\operatorname{sen} 0 + 4 \operatorname{sen} \frac{\pi}{4} + \operatorname{sen} \frac{\pi}{2}\right] = \frac{\pi}{12}(2\sqrt{2}+1) = 1.002279878$$

y

$$S\left(0, \frac{\pi}{4}\right) + S\left(\frac{\pi}{4}, \frac{\pi}{2}\right) = \frac{\pi/8}{3}\left[\operatorname{sen} 0 + 4 \operatorname{sen} \frac{\pi}{8} + 2 \operatorname{sen} \frac{\pi}{4} + 4 \operatorname{sen} \frac{3\pi}{8} + \operatorname{sen} \frac{\pi}{2}\right]$$
$$= 1.000134585.$$

Por lo que,

$$\left| S\left(0, \frac{\pi}{2}\right) - S\left(0, \frac{\pi}{4}\right) - S\left(\frac{\pi}{4}, \frac{\pi}{2}\right) \right| = |1.002279878 - 1.000134585| = 0.002145293.$$

El cálculo para el error obtenido al utilizar $S(a, (a+b)) + S((a+b), b)$ para aproximar $\int_a^b f(x)\,dx$ es, por consiguiente,

$$\frac{1}{15}\left| S\left(0, \frac{\pi}{2}\right) - S\left(0, \frac{\pi}{4}\right) - S\left(\frac{\pi}{4}, \frac{\pi}{2}\right) \right| = 0.000143020,$$

el cual se aproxima de cerca al error real

$$\left| \int_0^{\pi/2} \text{sen}\, x\, dx - 1.000134585 \right| = 0.000134585,$$

aunque D_x^4 sen $x =$ sen x varía significativamente en el intervalo $(0, \pi/2)$. ■

Cuando las aproximaciones en la desigualdad (4.38) difieren por más de 15ε, podemos aplicar la técnica de la regla de Simpson de manera individual a los subintervalos $[a, (a+b)/2]$ y $[(a+b)/2, b]$. A continuación, usamos el procedimiento de cálculo de error para determinar si la aproximación para la integral en cada subintervalo se encuentra dentro de una tolerancia de $\varepsilon/2$. En este caso, sumamos las aproximaciones para producir una aproximación de $\int_a^b f(x)\, dx$ dentro de la tolerancia ε.

Si la aproximación en uno de los subintervalos no se encuentra dentro de la tolerancia $\varepsilon/2$, entonces ese subintervalo se subdivide en sí mismo y el procedimiento se reaplica a los dos subintervalos para determinar si la aproximación en cada subintervalo es precisa dentro de $\varepsilon/4$. Este procedimiento de reducción a la mitad continúa hasta que cada parte está dentro de la tolerancia requerida.

Se pueden construir problemas para los que esta tolerancia nunca se cumplirá, pero normalmente la técnica es exitosa porque cada subdivisión aumenta la precisión de la aproximación en un factor de 16 mientras requiere un factor de precisión aumentado en sólo 2.

El algoritmo 4.3 describe detalladamente este procedimiento de cuadratura adaptable para la regla de Simpson, a pesar de que la implementación difiere ligeramente del análisis anterior. Por ejemplo, en el paso 1, la tolerancia se ha establecido en 10ε en lugar de en 15ε en la desigualdad (4.38). Esta cota se elige de manera conservadora para compensar el error en la suposición $f^{(4)}(\xi) \approx f^{(4)}(\tilde{\xi})$. En problemas donde se sabe que $f^{(4)}$ varía ampliamente, esta cota debería disminuir todavía más.

Es buena idea incluir un margen de seguridad cuando es imposible verificar las suposiciones de precisión.

El procedimiento listado en el algoritmo, primero aproxima la integral en el subintervalo situado más a la izquierda en una subdivisión. Esto requiere almacenamiento eficiente y recordar las evaluaciones funcionales que se han calculado antes para los nodos en los subintervalos situados a la mitad derecha. Los pasos 3, 4 y 5 contienen un procedimiento de apilamiento con un indicador para seguir los datos requeridos para calcular la aproximación en el subintervalo inmediatamente adyacente y a la derecha del subintervalo en el que se genera la aproximación. El método es más fácil de implementar por medio de lenguaje de programación recursivo.

ALGORITMO **4.3**

Cuadratura adaptable

Para aproximar la integral $I = \int_a^b f(x)\, dx$ dentro de una tolerancia determinada:

ENTRADA extremos a, b; tolerancia TOL; cota N para diferentes niveles.

SALIDA aproximación APP o mensaje N superado.

Paso 1 Haga $APP = 0$;
$i = 1$;
$TOL_i = 10\, TOL$;
$a_i = a$;
$h_i = (b-a)/2$;
$FA_i = f(a)$;
$FC_i = f(a + h_i)$;
$FB_i = f(b)$;
$S_i = h_i(FA_i + 4FC_i + FB_i)/3$; (*Aproximación a partir del método de Simpson para intervalo completo*.)
$L_i = 1$.

Paso 2 Si $i > 0$ haga los pasos 3–5.

Paso 3 Haga $FD = f(a_i + h_i/2)$;
$FE = f(a_i + 3h_i/2)$;
$S1 = h_i(FA_i + 4FD + FC_i)/6$; (*Aproximaciones para el método de Simpson para mitades de subintervalos.*)
$S2 = h_i(FC_i + 4FE + FB_i)/6$;
$v_1 = a_i$; (*Guardar datos en este nivel.*)
$v_2 = FA_i$;
$v_3 = FC_i$;
$v_4 = FB_i$;
$v_5 = h_i$;
$v_6 = TOL_i$;
$v_7 = S_i$;
$v_8 = L_i$.

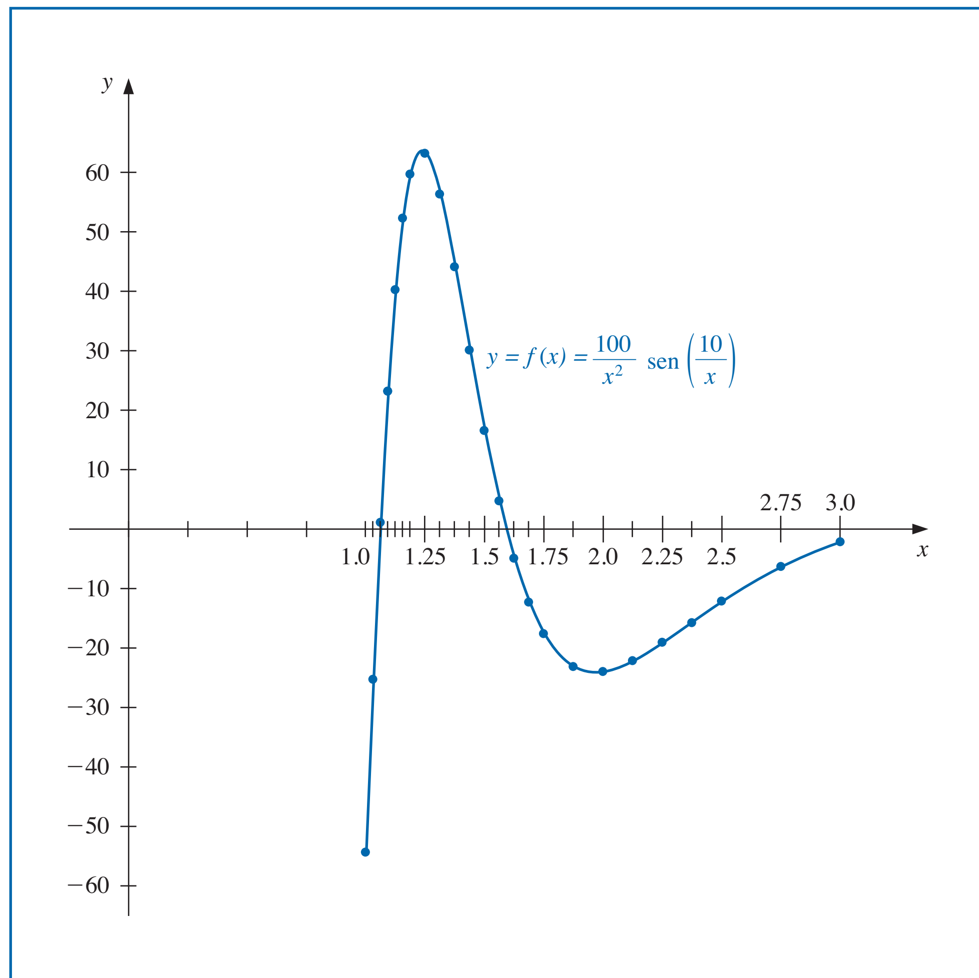
Paso 4 Haga $i = i - 1$. (*Borrar el nivel.*)

Paso 5 Si $|S1 + S2 - v_7| < v_6$
entonces determine $APP = APP + (S1 + S2)$
también
si ($v_8 \geq N$)
entonces
SALIDA ('NIVEL EXCEDIDO'); (*Falla del procedimiento.*)
PARE.
también (*Añadir un nivel.*)
haga $i = i + 1$; (*Datos para subintervalo mitad derecha.*)
$a_i = v_1 + v_5$;
$FA_i = v_3$;
$FC_i = FE$;
$FB_i = v_4$;
$h_i = v_5/2$;
$TOL_i = v_6/2$;
$S_i = S2$;
$L_i = v_8 + 1$;
haga $i = i + 1$; (*Datos para subintervalo mitad izquierda.*)
$a_i = v_1$;
$FA_i = v_2$;
$FC_i = FD$;
$FB_i = v_3$;
$h_i = h_{i-1}$;
$TOL_i = TOL_{i-1}$;
$S_i = S1$;
$L_i = L_{i-1}$.

Paso 6 SALIDA (APP); (*APP se aproxima a I dentro de TOL.*)
PARE. ■

Ilustración La gráfica de la función $f(x) = (100/x^2)\,\text{sen}(10/x)$ para x en [1, 3] se muestra en la figura 4.13. Por medio del algoritmo de cuadratura adaptable 4.3 con tolerancia 10^{-4} para aproximar $\int_1^3 f(x)\,dx$ produce -1.426014, un resultado preciso dentro de 1.1×10^{-5}. La aproximación requería que la regla de Simpson con $n = 4$ se realice sobre los 23 subintervalos cuyos extremos se muestran en el eje horizontal en la figura 4.14. El número total de evaluaciones funcionales requerido para esta aproximación es 93.

Figura 4.14



El valor más grande de h para el que la regla compuesta de Simpson estándar provee precisión dentro de 10^{-4} es $h = 1/88$. Esta aplicación requiere 177 evaluaciones de función, aproximadamente dos veces más que la cuadratura adaptable. ■

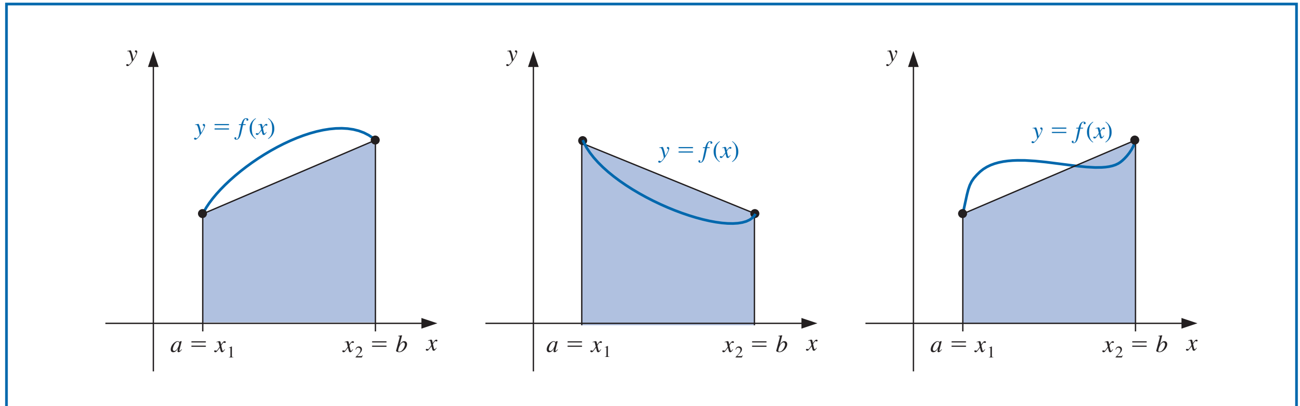
La sección Conjunto de ejercicios 4.6 está disponible en línea: Encuentre la ruta de acceso en las páginas preliminares.

4.7 Cuadratura gaussiana

Las fórmulas de Newton-Cotes en la sección 4.3 se dedujeron al integrar polinomios de interpolación. El término de error en el polinomio interpolante de grado n implica la derivada $(n + 1)$ de la función que se va a aproximar, por lo que la fórmula Newton-Cotes es exacta al aproximar la integral de cualquier polinomio de grado menor o igual a n.

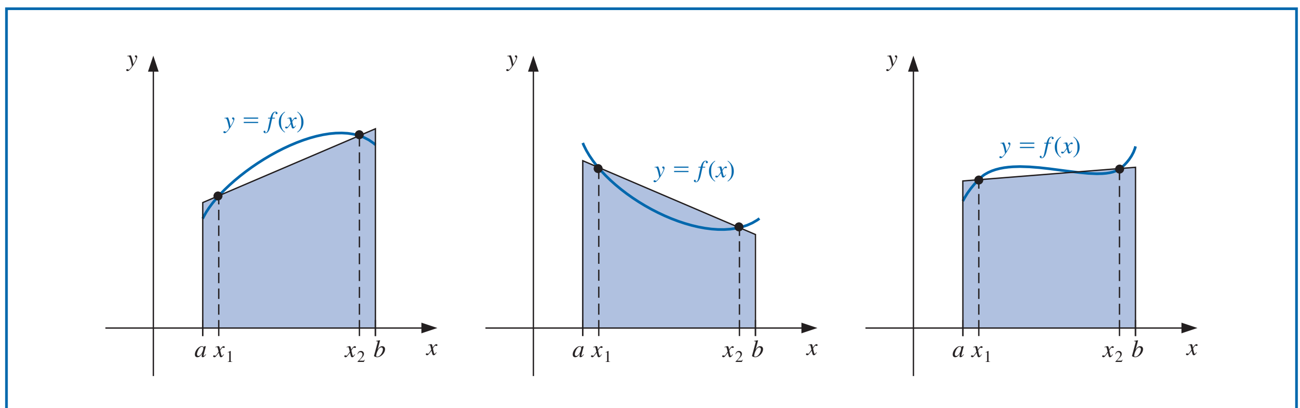
Todas las fórmulas de Newton-Cotes utilizan valores de la función en puntos igualmente espaciados. Esta restricción es conveniente cuando se combinan las fórmulas para formar las reglas compuestas que consideramos en la sección 4.4, pero puede disminuir significativamente la precisión de la aproximación. Considere, por ejemplo, la regla trapezoidal aplicada para determinar las integrales de las funciones cuyas gráficas se muestran en la figura 4.15.

Figura 4.15



La regla trapezoidal aproxima la integral de la función al integrar la función lineal que une los extremos de la gráfica de la función. Pero probablemente no es la mejor línea para aproximar la integral. En muchos casos, las líneas como las mostradas en la figura 4.16 probablemente proporcionarían mucho mejores aproximaciones.

Figura 4.16



En la cuadratura Gaussiana, los puntos para evaluación se seleccionan de manera óptima en lugar de nodos igualmente espaciados. Los nodos $x_1, x_2, \ldots, x_n$ en el intervalo $[a, b]$, y los coeficientes $c_1, c_2, \ldots, c_n$ se seleccionan para minimizar el error esperado obtenido en la aproximación

$$\int_a^b f(x)\,dx \approx \sum_{i=1}^{n} c_i f(x_i).$$

Gauss demostró su método de integración numérica eficiente en un documento presentado ante la Göttingen Society en 1814. Él permitió que los nodos, así como los coeficientes de las evaluaciones de función, fueran parámetros en la fórmula de suma y encontró la colocación óptima de los nodos. Goldstine [Golds], pp. 224–232, tiene una descripción interesante de este desarrollo.

Para medir esta precisión, suponemos que la mejor elección de estos valores produce el resultado exacto para la clase de polinomios de mayor grado, es decir, la selección que da el grado más alto de precisión.

Los coeficientes $c_1, c_2, \ldots, c_n$ en la fórmula de aproximación son arbitrarios y los nodos $x_1, x_2, \ldots, x_n$ están restringidos solamente por el hecho de que deben encontrarse en $[a, b]$, el intervalo de integración. Esto nos da $2n$ parámetros a elegir. Si los coeficientes de un polinomio se consideran parámetros, la clase de polinomios de grado por lo menos $2n - 1$ también contiene $2n$ parámetros. Esto, entonces, es la mayor clase de polinomios para los que es razonable esperar que una fórmula sea exacta. Con la elección adecuada de valores y constantes, se puede obtener la precisión en este conjunto.

Para ilustrar el procedimiento para elegir los parámetros adecuados, mostraremos cómo seleccionar los coeficientes y nodos cuando $n = 2$ y el intervalo de integración es $[-1, 1]$. Entonces, analizaremos la situación más general para una selección arbitraria de nodos y coeficientes y mostraremos cómo se modifica la técnica al integrar sobre un intervalo arbitrario.

Suponga que queremos determinar c_1, c_2, x_1, y x_2 de tal forma que la fórmula de integración

$$\int_{-1}^{1} f(x)\,dx \approx c_1 f(x_1) + c_2 f(x_2)$$

da el resultado exacto siempre que $f(x)$ es un polinomio de grado $2(2) - 1 = 3$ o menor, es decir, cuando

$$f(x) = a_0 + a_1 x + a_2 x^2 + a_3 x^3,$$

para algún conjunto de constantes a_0, a_1, a_2, y a_3. Puesto que

$$\int (a_0 + a_1 x + a_2 x^2 + a_3 x^3)\,dx = a_0 \int 1\,dx + a_1 \int x\,dx + a_2 \int x^2\,dx + a_3 \int x^3\,dx,$$

esto es equivalente a mostrar que la fórmula da resultados exactos cuando $f(x)$ es 1, x, x^2, y x^3. Por lo tanto, necesitamos c_1, c_2, x_1, y x_2, de tal forma que

$$c_1 \cdot 1 + c_2 \cdot 1 = \int_{-1}^{1} 1\,dx = 2, \qquad c_1 \cdot x_1 + c_2 \cdot x_2 = \int_{-1}^{1} x\,dx = 0,$$

$$c_1 \cdot x_1^2 + c_2 \cdot x_2^2 = \int_{-1}^{1} x^2\,dx = \frac{2}{3}, \quad \text{y} \quad c_1 \cdot x_1^3 + c_2 \cdot x_2^3 = \int_{-1}^{1} x^3\,dx = 0.$$

Un poco de álgebra muestra que este sistema de ecuaciones tiene una solución única

$$c_1 = 1, \quad c_2 = 1, \quad x_1 = -\frac{\sqrt{3}}{3}, \quad \text{y} \quad x_2 = \frac{\sqrt{3}}{3},$$

lo cual da la fórmula de aproximación

$$\int_{-1}^{1} f(x)\,dx \approx f\left(\frac{-\sqrt{3}}{3}\right) + f\left(\frac{\sqrt{3}}{3}\right). \tag{4.40}$$

Esta fórmula tiene grado de precisión tres; es decir, produce el resultado exacto para cada polinomio de grado tres o menor.

Polinomios de Legendre

La técnica que hemos descrito se puede utilizar para determinar los nodos y coeficientes para las fórmulas que dan resultados exactos para polinomios de grado superior, pero un método alterno los obtiene con mayor facilidad. En las secciones 8.2 y 8.3, consideraremos diferentes conjuntos de polinomios ortogonales, funciones que tienen la propiedad de que una integral definida del producto de cualquiera de dos de ellas es 0. El conjunto relevante para nuestro problema son los polinomios de Legendre, un conjunto $\{P_0(x), P_1(x), \ldots, P_n(x), \ldots,\}$ con las siguientes propiedades:

(1) Para cada n, $P_n(x)$ es un polinomio mónico de grado n.

(2) $\displaystyle\int_{-1}^{1} P(x)P_n(x)\,dx = 0$ siempre que $P(x)$ sea un polinomio de grado menor a n.

Recuerde que los polinomios *mónicos* tienen un coeficiente principal de 1.

Los primeros polinomios de Legendre son

$$P_0(x) = 1, \quad P_1(x) = x, \quad P_2(x) = x^2 - \frac{1}{3},$$
$$P_3(x) = x^3 - \frac{3}{5}x, \quad \text{y} \quad P_4(x) = x^4 - \frac{6}{7}x^2 + \frac{3}{35}.$$

Adrien-Marie Legendre (1752–1833) introdujo este conjunto de polinomios en 1785. Tuvo numerosas disputas prioritarias con Gauss, principalmente debido a la falla de Gauss al publicar muchos de sus resultados originales mucho después de que los había descubierto.

Las raíces de estos polinomios son distintas, se encuentran en el intervalo $(-1, 1)$, tienen una simetría respecto al origen y, más importante, son la elección correcta para determinar los parámetros que nos proporcionan los nodos y coeficientes para nuestro método de cuadratura.

Los nodos $x_1, x_2, \ldots, x_n$ necesarios para producir una fórmula de aproximación integral que proporcione resultados exactos para cualquier polinomio de grado menor a $2n$ son las raíces del polinomio de Legendre de enésimo grado. Esto se establece mediante el siguiente resultado.

Teorema 4.7 Suponga que $x_1, x_2, \ldots, x_n$ son las raíces del n-ésimo polinomio de Legendre $P_n(x)$ y que para cada $i = 1, 2, \ldots, n$, los números c_i están definidos por

$$c_i = \int_{-1}^{1} \prod_{\substack{j=1 \\ j\neq i}}^{n} \frac{x - x_j}{x_i - x_j}\, dx.$$

Si $P(x)$ es cualquier polinomio de grado menor a $2n$, entonces

$$\int_{-1}^{1} P(x)\, dx = \sum_{i=1}^{n} c_i P(x_i).$$

Demostración Consideramos primero la situación para un polinomio $P(x)$ de grado menor a n. Reescriba $P(x)$ en términos de los polinomios de coeficientes de Lagrange $(n-1)$-ésimos con nodos en las raíces del n-ésimo polinomio de Legendre $P_n(x)$. El término de error para esta representación implica la n-ésima derivada de $P(x)$. Puesto que $P(x)$ es de grado menor a n, la n-ésima derivada de $P(x)$ es 0 y esta representación es exacta. Por lo tanto,

$$P(x) = \sum_{i=1}^{n} P(x_i)L_i(x) = \sum_{i=1}^{n} \prod_{\substack{j=1 \\ j\neq i}}^{n} \frac{x - x_j}{x_i - x_j} P(x_i)$$

y

$$\int_{-1}^{1} P(x)\, dx = \int_{-1}^{1} \left[\sum_{i=1}^{n} \prod_{\substack{j=1 \\ j\neq i}}^{n} \frac{x - x_j}{x_i - x_j} P(x_i) \right] dx$$
$$= \sum_{i=1}^{n} \left[\int_{-1}^{1} \prod_{\substack{j=1 \\ j\neq i}}^{n} \frac{x - x_j}{x_i - x_j}\, dx \right] P(x_i) = \sum_{i=1}^{n} c_i P(x_i).$$

Por lo tanto, el resultado es verdad para polinomios de grado menor a n.

Ahora considere un polinomio $P(x)$ de grado por lo menos n pero menor a $2n$. Divida $P(x)$ entre el n-ésimo polinomio de Legendre $P_n(x)$. Esto proporciona dos polinomios $Q(x)$ y $R(x)$, cada uno de grado menor a n, con

$$P(x) = Q(x)P_n(x) + R(x).$$

Observe que x_i es una raíz de $P_n(x)$ para cada $i = 1, 2, \ldots, n$, por lo que tenemos

$$P(x_i) = Q(x_i)P_n(x_i) + R(x_i) = R(x_i).$$

Ahora, recurrimos a la potencia única de los polinomios de Legendre. En primer lugar, el grado del polinomio $Q(x)$ es menor a n, por lo que (mediante la propiedad (2) de Legendre),

$$\int_{-1}^{1} Q(x)P_n(x)\,dx = 0.$$

A continuación, puesto que $R(x)$ es un polinomio de grado menor a n, el argumento de apertura implica que

$$\int_{-1}^{1} R(x)\,dx = \sum_{i=1}^{n} c_i R(x_i).$$

Al juntar estos hechos, verificamos que la fórmula es exacta para el polinomio $P(x)$:

$$\int_{-1}^{1} P(x)\,dx = \int_{-1}^{1} [Q(x)P_n(x) + R(x)]\,dx = \int_{-1}^{1} R(x)\,dx = \sum_{i=1}^{n} c_i R(x_i) = \sum_{i=1}^{n} c_i P(x_i).$$

■

Las constantes c_i necesarias para la regla de cuadratura se pueden generar a partir de la ecuación en el teorema 4.7, pero tanto las constantes como las raíces de los polinomios de Legendre se tabulan ampliamente. La tabla 4.12 enumera estos valores para $n = 2, 3, 4$ y 5.

Tabla 4.12

n	Raíces $r_{n,i}$	Coeficientes $c_{n,i}$
2	0.5773502692	1.0000000000
	−0.5773502692	1.0000000000
3	0.7745966692	0.5555555556
	0.0000000000	0.8888888889
	−0.7745966692	0.5555555556
4	0.8611363116	0.3478548451
	0.3399810436	0.6521451549
	−0.3399810436	0.6521451549
	−0.8611363116	0.3478548451
5	0.9061798459	0.2369268850
	0.5384693101	0.4786286705
	0.0000000000	0.5688888889
	−0.5384693101	0.4786286705
	−0.9061798459	0.2369268850

Ejemplo 1 Aproxime $\int_{-1}^{1} e^x \cos x\,dx$ mediante cuadratura gaussiana con $n = 3$.

Solución Las entradas en la tabla 4.12 nos dan

$$\int_{-1}^{1} e^x \cos x\,dx \approx 0.\overline{5}e^{0.774596692}\cos 0.774596692$$

$$+ 0.\overline{8}\cos 0 + 0.\overline{5}e^{-0.774596692}\cos(-0.774596692)$$

$$= 1.9333904.$$

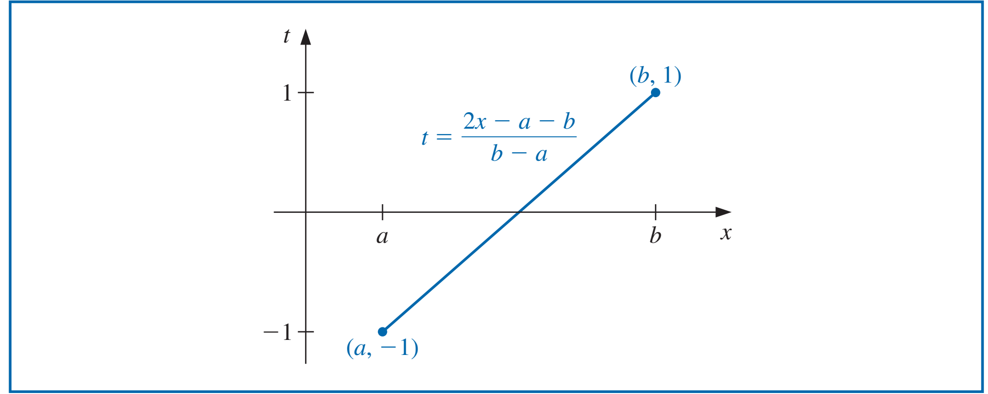
La integración por partes se puede utilizar para mostrar que el valor verdadero de la integral es 1.9334214, por lo que el error absoluto es menor a 3.2×10^{-5}. ■

Cuadratura gaussiana en intervalos arbitrarios

Una integral $\int_a^b f(x)\,dx$ sobre un intervalo arbitrario $[a, b]$ se puede transformar en una integral sobre $[-1, 1]$ al utilizar el cambio de variables (véase la figura 4.17):

$$t = \frac{2x - a - b}{b - a} \Longleftrightarrow x = \frac{1}{2}[(b - a)t + a + b].$$

Figura 4.17



Esto permite aplicar la cuadratura gaussiana a cualquier intervalo $[a, b]$ porque

$$\int_a^b f(x)\,dx = \int_{-1}^{1} f\left(\frac{(b - a)t + (b + a)}{2}\right)\frac{(b - a)}{2}\,dt. \tag{4.41}$$

Ejemplo 2 Considere la integral $\displaystyle\int_1^3 x^6 - x^2 \operatorname{sen}(2x)\,dx = 317.3442466.$

a) Compare los resultados de la fórmula cerrada de Newton-Cotes con $n = 1$, la fórmula abierta de Newton-Cotes con $n = 1$, y la cuadratura gaussiana cuando $n = 2$.

b) Compare los resultados de la fórmula cerrada de Newton-Cotes con $n = 2$, la fórmula abierta de Newton-Cotes con $n = 2$, y la cuadratura gaussiana cuando $n = 3$.

Solución a) Cada una de las fórmulas en esta parte requiere dos evaluaciones de la función $f(x) = x^6 - x^2 \operatorname{sen}(2x)$. Las aproximaciones de Newton-Cotes son

$$\textbf{Cerrada } \boldsymbol{n = 1}: \quad \frac{2}{2}[f(1) + f(3)] = 731.6054420;$$

$$\textbf{Abierta } \boldsymbol{n = 1}: \quad \frac{3(2/3)}{2}[f(5/3) + f(7/3)] = 188.7856682.$$

La cuadratura gaussiana aplicada a este problema requiere que la integral primero se transforme en un problema cuyo intervalo de integración es $[-1, 1]$. Por medio de la ecuación (4.41) obtenemos

$$\int_1^3 x^6 - x^2 \operatorname{sen}(2x)\,dx = \int_{-1}^{1} (t + 2)^6 - (t + 2)^2 \operatorname{sen}(2(t + 2))\,dt.$$

Entonces, la cuadratura con $n = 2$ da

$$\int_1^3 x^6 - x^2 \operatorname{sen}(2x)\,dx \approx f(-0.5773502692 + 2) + f(0.5773502692 + 2)$$
$$= 306.8199344.$$

b) Cada una de las fórmulas en esta parte requiere tres evaluaciones de función. Las aproximaciones de Newton-Cotes son

$$\textbf{Cerrada } \boldsymbol{n} = \mathbf{2}: \quad \frac{1}{3}[f(1) + 4f(2) + f(3)] = 333.2380940;$$

$$\textbf{Abierta } \boldsymbol{n} = \mathbf{2}: \quad \frac{4(1/2)}{3}[2f(1.5) - f(2) + 2f(2.5)] = 303.5912023.$$

La cuadratura gaussiana con $n = 3$, una vez que la transformación se ha realizado, obtenemos

$$\int_1^3 x^6 - x^2 \operatorname{sen}(2x)\, dx \approx 0.\overline{5} f(-0.7745966692 + 2)$$
$$+ 0.\overline{8} f(2) + 0.\overline{5} f(-0.7745966692 + 2) = 317.2641516.$$

Los resultados de la cuadratura gaussiana son claramente superiores en cada instancia. ■

La sección Conjunto de ejercicios 4.7 está disponible en línea. Encuentre la ruta de acceso en las páginas preliminares.

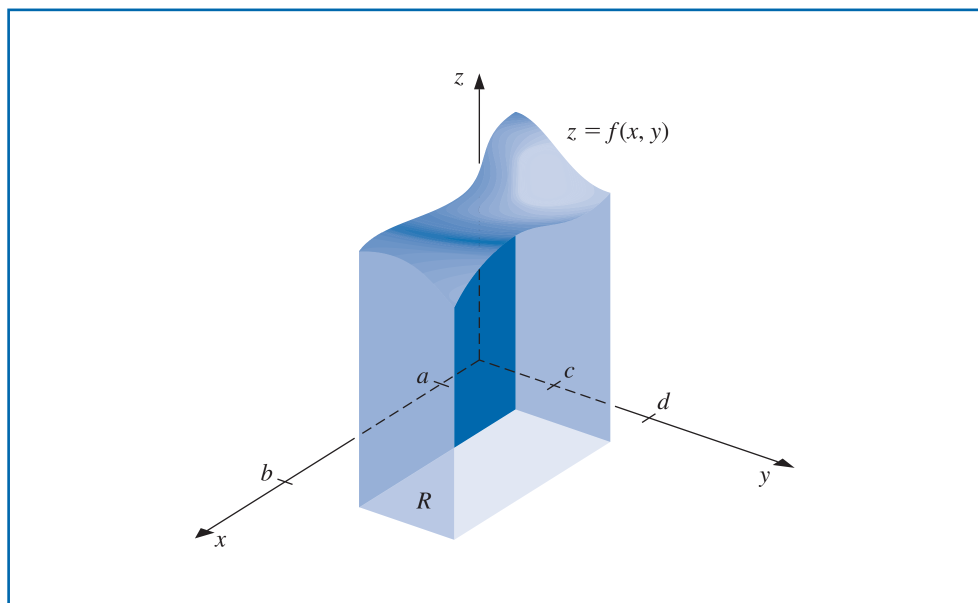
4.8 Integrales múltiples

Las técnicas analizadas en las secciones previas se pueden modificar para utilizarse en la aproximación de integrales múltiples. Considere la integral doble

$$\iint_R f(x, y)\, dA,$$

cuando $R = \{(x, y) \mid a \leq x \leq b, c \leq y \leq d\}$, para algunas constantes a, b, c y d, es una región rectangular en el plano. (Véase la figura 4.18.)

Figura 4.18



La siguiente ilustración muestra la manera en la que la regla compuesta trapezoidal por medio de dos subintervalos en cada dirección coordinada se aplicaría a esta integral.

Ilustración Al escribir la integral doble como una integral iterada obtenemos

$$\iint_R f(x, y)\, dA = \int_a^b \left(\int_c^d f(x, y)\, dy \right) dx.$$

Para simplificar la notación, si $k = (d-c)/2$ y $h = (b-a)/2$. Aplique la regla compuesta trapezoidal a la integral interior para obtener

$$\int_c^d f(x, y)\, dy \approx \frac{k}{2} \left[f(x, c) + f(x, d) + 2f\left(x, \frac{c+d}{2}\right) \right].$$

Esta aproximación es de orden $O\left((d-c)^3\right)$. A continuación, aplique la regla compuesta trapezoidal nuevamente para aproximar la integral de esta función de x:

$$\int_a^b \left(\int_c^d f(x, y)\, dy \right) dx \approx \int_a^b \left(\frac{d-c}{4} \right) \left[f(x, c) + 2f\left(x, \frac{c+d}{2}\right) + f(x, d) \right] dx$$

$$= \frac{b-a}{4} \left(\frac{d-c}{4} \right) \left[f(a, c) + 2f\left(a, \frac{c+d}{2}\right) + f(a, d) \right]$$

$$+ \frac{b-a}{4} \left(2 \left(\frac{d-c}{4} \right) \left[f\left(\frac{a+b}{2}, c \right) + 2f\left(\frac{a+b}{2}, \frac{c+d}{2} \right) + f\left(\frac{a+b}{2}, d \right) \right] \right)$$

$$+ \frac{b-a}{4} \left(\frac{d-c}{4} \right) \left[f(b, c) + 2f\left(b, \frac{c+d}{2}\right) + f(b, d) \right]$$

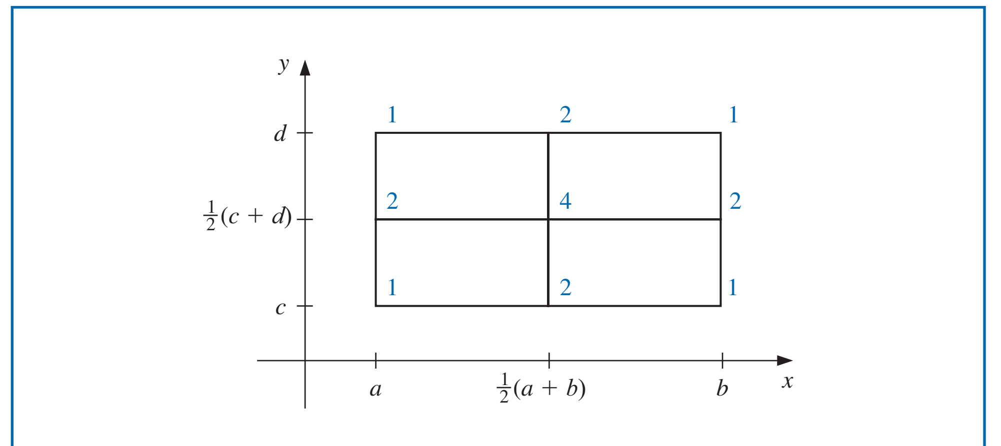
$$= \frac{(b-a)(d-c)}{16} \left[f(a, c) + f(a, d) + f(b, c) + f(b, d) \right.$$

$$+ 2\left(f\left(\frac{a+b}{2}, c \right) + f\left(\frac{a+b}{2}, d \right) + f\left(a, \frac{c+d}{2} \right) + f\left(b, \frac{c+d}{2} \right) \right)$$

$$\left. + 4f\left(\frac{a+b}{2}, \frac{c+d}{2} \right) \right]$$

Esta aproximación es de orden $O\left((b-a)(d-c)\left[(b-a)^2 + (d-c)^2\right]\right)$. La figura 4.19 muestra una cuadrícula con el número de evaluaciones funcionales en cada uno de los nodos utilizados en la aproximación. ■

Figura 4.19



Como muestra la ilustración, el procedimiento es bastante sencillo. Sin embargo, el número de evaluaciones de función crece con el cuadrado del número requerido para una sola integral. En una situación práctica, no esperaríamos utilizar un método tan básico como la regla compuesta trapezoidal con $n = 2$. En cambio, usaremos la regla compuesta de Simpson que es más apropiada para ilustrar la técnica de aproximación general, a pesar de que se podría usar cualquier otra fórmula en su lugar.

Para aplicar la regla compuesta de Simpson, dividimos la región R al subdividir $[a, b]$ y $[c, d]$ en un número par de subintervalos. Para simplificar la notación, seleccionamos enteros pares n y m y subdividimos $[a, b]$ y $[c, d]$ con puntos de malla igualmente espaciados $x_0, x_1, \ldots, x_n$ y $y_0, y_1, \ldots, y_m$, respectivamente. Estas subdivisiones determinan tamaños de pasos $h = (b - a)/n$ y $k = (d - c)/m$. Al escribir la integral doble como la integral iterada

$$\iint_R f(x, y)\, dA = \int_a^b \left(\int_c^d f(x, y)\, dy \right) dx,$$

primero utilizamos la regla compuesta de Simpson para aproximar

$$\int_c^d f(x, y)\, dy,$$

tomando x como una constante.

Si $y_j = c + jk$, para cada $j = 0, 1, \ldots, m$. Entonces

$$\int_c^d f(x, y)\, dy = \frac{k}{3}\left[f(x, y_0) + 2\sum_{j=1}^{(m/2)-1} f(x, y_{2j}) + 4\sum_{j=1}^{m/2} f(x, y_{2j-1}) + f(x, y_m) \right] - \frac{(d-c)k^4}{180}\frac{\partial^4 f(x, \mu)}{\partial y^4},$$

para algunas μ en (c, d). Por lo tanto,

$$\begin{aligned}\int_a^b \int_c^d f(x, y)\, dy\, dx = \frac{k}{3}\Bigg[&\int_a^b f(x, y_0)\, dx + 2\sum_{j=1}^{(m/2)-1} \int_a^b f(x, y_{2j})\, dx \\ &+ 4\sum_{j=1}^{m/2} \int_a^b f(x, y_{2j-1})\, dx + \int_a^b f(x, y_m)\, dx \Bigg] \\ &- \frac{(d-c)k^4}{180} \int_a^b \frac{\partial^4 f(x, \mu)}{\partial y^4}\, dx.\end{aligned}$$

Ahora, utilizamos la regla compuesta de Simpson para las integrales en esta ecuación. Si $x_i = a + ih$, para cada $i = 0, 1, \ldots, n$. Entonces, para cada $j = 0, 1, \ldots, m$ tenemos

$$\int_a^b f(x, y_j)\, dx = \frac{h}{3}\left[f(x_0, y_j) + 2\sum_{i=1}^{(n/2)-1} f(x_{2i}, y_j) + 4\sum_{i=1}^{n/2} f(x_{2i-1}, y_j) + f(x_n, y_j) \right] - \frac{(b-a)h^4}{180}\frac{\partial^4 f}{\partial x^4}(\xi_j, y_j),$$

para algunos ξ_j en (a, b). La aproximación resultante tiene la forma

$$\int_a^b \int_c^d f(x, y)\,dy\,dx \approx \frac{hk}{9}\Bigg\{\Bigg[f(x_0, y_0) + 2\sum_{i=1}^{(n/2)-1} f(x_{2i}, y_0) + 4\sum_{i=1}^{n/2} f(x_{2i-1}, y_0) + f(x_n, y_0)\Bigg]$$
$$+ 2\Bigg[\sum_{j=1}^{(m/2)-1} f(x_0, y_{2j}) + 2\sum_{j=1}^{(m/2)-1}\sum_{i=1}^{(n/2)-1} f(x_{2i}, y_{2j}) + 4\sum_{j=1}^{(m/2)-1}\sum_{i=1}^{n/2} f(x_{2i-1}, y_{2j}) + \sum_{j=1}^{(m/2)-1} f(x_n, y_{2j})\Bigg]$$
$$+ 4\Bigg[\sum_{j=1}^{m/2} f(x_0, y_{2j-1}) + 2\sum_{j=1}^{m/2}\sum_{i=1}^{(n/2)-1} f(x_{2i}, y_{2j-1}) + 4\sum_{j=1}^{m/2}\sum_{i=1}^{n/2} f(x_{2i-1}, y_{2j-1}) + \sum_{j=1}^{m/2} f(x_n, y_{2j-1})\Bigg]$$
$$+ \Bigg[f(x_0, y_m) + 2\sum_{i=1}^{(n/2)-1} f(x_{2i}, y_m) + 4\sum_{i=1}^{n/2} f(x_{2i-1}, y_m) + f(x_n, y_m)\Bigg]\Bigg\}.$$

El término de error E está dado por

$$E = \frac{-k(b-a)h^4}{540}\Bigg[\frac{\partial^4 f(\xi_0, y_0)}{\partial x^4} + 2\sum_{j=1}^{(m/2)-1}\frac{\partial^4 f(\xi_{2j}, y_{2j})}{\partial x^4} + 4\sum_{j=1}^{m/2}\frac{\partial^4 f(\xi_{2j-1}, y_{2j-1})}{\partial x^4} + \frac{\partial^4 f(\xi_m, y_m)}{\partial x^4}\Bigg] - \frac{(d-c)k^4}{180}\int_a^b \frac{\partial^4 f(x, \mu)}{\partial y^4}\,dx.$$

Si $\partial^4 f/\partial x^4$ es continua, el teorema del valor intermedio 1.11 se puede aplicar repetidamente para mostrar que la evaluación de las derivadas parciales respecto a x se pueden reemplazar con un valor común y que

$$E = \frac{-k(b-a)h^4}{540}\left[3m\frac{\partial^4 f}{\partial x^4}(\overline{\eta}, \overline{\mu})\right] - \frac{(d-c)k^4}{180}\int_a^b \frac{\partial^4 f(x, \mu)}{\partial y^4}\,dx,$$

para algunos $(\overline{\eta}, \overline{\mu})$ en R. Si $\partial^4 f/\partial y^4$ también es continua, el teorema de valor medio para integrales implica que

$$\int_a^b \frac{\partial^4 f(x, \mu)}{\partial y^4}\,dx = (b-a)\frac{\partial^4 f}{\partial y^4}(\hat{\eta}, \hat{\mu}),$$

para algunos $(\hat{\eta}, \hat{\mu})$ en R. Puesto que $m = (d-c)/k$, el término de error tiene la forma

$$E = \frac{-k(b-a)h^4}{540}\left[3m\frac{\partial^4 f}{\partial x^4}(\overline{\eta}, \overline{\mu})\right] - \frac{(d-c)(b-a)}{180}k^4\frac{\partial^4 f}{\partial y^4}(\hat{\eta}, \hat{\mu}),$$

lo cual se simplifica en

$$E = -\frac{(d-c)(b-a)}{180}\left[h^4\frac{\partial^4 f}{\partial x^4}(\overline{\eta}, \overline{\mu}) + k^4\frac{\partial^4 f}{\partial y^4}(\hat{\eta}, \hat{\mu})\right],$$

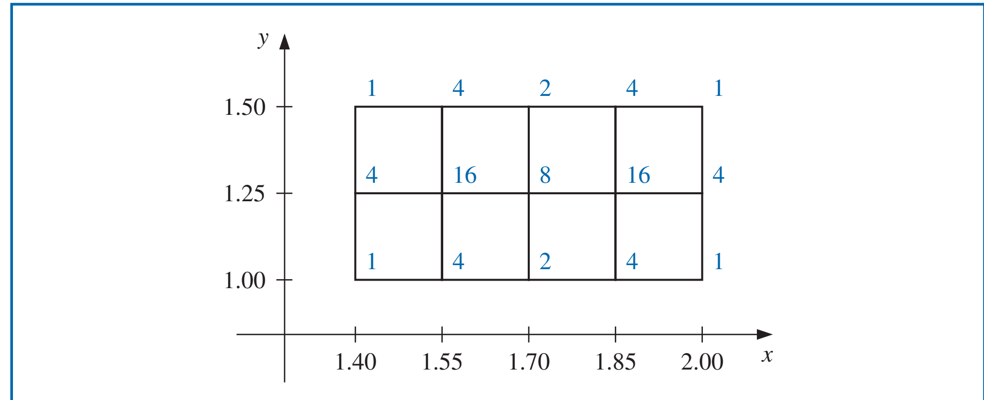
para algunos $(\overline{\eta}, \overline{\mu})$ y $(\hat{\eta}, \hat{\mu})$ en R.

Ejemplo 1 Utilice la regla compuesta de Simpson con $n = 4$ y $m = 2$ para aproximar

$$\int_{1.4}^{2.0}\int_{1.0}^{1.5} \ln(x+2y)\,dy\,dx.$$

Solución Los tamaños de paso para esta aplicación son $h = (2.0-1.4)/4 = 0.15$ y $k = (1.5-1.0)/2 = 0.25$. La región de integración R se muestra en la figura 4.20, junto con los nodos (x_i, y_j), donde $i = 0, 1, 2, 3, 4$ y $j = 0, 1, 2$. También muestra que los coeficientes $w_{i,j}$ de $f(x_i, y_i) = \ln(x_i + 2y_i)$ en la suma que da la aproximación de la regla compuesta de Simpson para la integral.

Figura 4.20



La aproximación es

$$\int_{1.4}^{2.0}\int_{1.0}^{1.5} \ln(x+2y)\,dy\,dx \approx \frac{(0.15)(0.25)}{9}\sum_{i=0}^{4}\sum_{j=0}^{2} w_{i,j}\ln(x_i + 2y_j)$$
$$= 0.4295524387.$$

tenemos

$$\frac{\partial^4 f}{\partial x^4}(x, y) = \frac{-6}{(x+2y)^4} \quad \text{y} \quad \frac{\partial^4 f}{\partial y^4}(x, y) = \frac{-96}{(x+2y)^4},$$

y los valores máximos de los valores absolutos de estas derivadas parciales se presentan en R cuando $x = 1.4$ y $y = 1.0$. Por lo que el error está acotado por

$$|E| \le \frac{(0.5)(0.6)}{180}\left[(0.15)^4 \max_{(x,y) \text{in} R}\frac{6}{(x+2y)^4} + (0.25)^4 \max_{(x,y) \text{in} R}\frac{96}{(x+2y)^4}\right] \le 4.72 \times 10^{-6}.$$

El valor real de la integral para 10 lugares decimales es

$$\int_{1.4}^{2.0}\int_{1.0}^{1.5} \ln(x+2y)\,dy\,dx = 0.4295545265,$$

por lo que la aproximación es precisa dentro de 2.1×10^{-6}. ■

Las mismas técnicas se pueden aplicar para la aproximación de las integrales triples así como integrales superiores para las funciones de más de tres variables. El número de evaluaciones requeridas para la aproximación es el producto del número de evaluaciones requeridas cuando se aplica el método a cada variable.

Cuadratura gaussiana para aproximación de integral doble

Para reducir el número de evaluaciones funcionales, se pueden incluir métodos más eficientes, como la cuadratura gaussiana, la integración de Romberg o la cuadratura adaptable, en lugar de las fórmulas de Newton-Cotes. El siguiente ejemplo ilustra el uso de cuadratura gaussiana para la integral considerada en el ejemplo 1.

Ejemplo 2 Utilice la cuadratura gaussiana con $n = 3$ en ambas dimensiones para aproximar la integral

$$\int_{1.4}^{2.0}\int_{1.0}^{1.5} \ln(x+2y)\,dy\,dx.$$

Solución Antes de emplear la cuadratura gaussiana para aproximar esta integral, necesitamos transformar la región de integración

$$R = \{\,(x, y) \mid 1.4 \le x \le 2.0,\, 1.0 \le y \le 1.5\,\} \quad \text{en}$$

$$\hat{R} = \{\,(u, v) \mid -1 \le u \le 1, -1 \le v \le 1\,\}.$$

Las transformaciones lineales que cumplen esto son

$$u = \frac{1}{2.0-1.4}(2x-1.4-2.0) \quad \text{y} \quad v = \frac{1}{1.5-1.0}(2y-1.0-1.5),$$

o, de manera equivalente $x = 0.3u + 1.7$ y $y = 0.25v + 1.25$. Al emplear este cambio de variables obtenemos una integral en la que se puede aplicar la cuadratura gaussiana:

$$\int_{1.4}^{2.0}\int_{1.0}^{1.5} \ln(x+2y)\,dy\,dx = 0.075\int_{-1}^{1}\int_{-1}^{1} \ln(0.3u+0.5v+4.2)\,dv\,du.$$

La fórmula de cuadratura gaussiana para $n = 3$ tanto en u como en v requiere que utilicemos los nodos

$$u_1 = v_1 = r_{3,2} = 0, \quad u_0 = v_0 = r_{3,1} = -0.7745966692,$$

y

$$u_2 = v_2 = r_{3,3} = 0.7745966692.$$

Los pesos asociados con $c_{3,2} = 0.\overline{8}$ y $c_{3,1} = c_{3,3} = 0.\overline{5}$. (Estos se muestran en la tabla 4.12 en la página 232.) La aproximación resultante es

$$\int_{1.4}^{2.0}\int_{1.0}^{1.5} \ln(x+2y)\,dy\,dx \approx 0.075\sum_{i=1}^{3}\sum_{j=1}^{3} c_{3,i}c_{3,j}\ln(0.3r_{3,i}+0.5r_{3,j}+4.2)$$

$$= 0.4295545313.$$

A pesar de que este resultado sólo requiere nueve evaluaciones funcionales en comparación con 15 para la regla compuesta de Simpson considerada en el ejemplo 1, éste es preciso dentro de 4.8×10^{-9}, en comparación con la precisión 2.1×10^{-6} en el ejemplo 1. ■

Regiones no rectangulares

El uso de métodos de aproximación para integrales dobles no está limitado a integrales con regiones rectangulares de integración. Las técnicas analizadas previamente se pueden modificar para aproximar las integrales dobles de la forma

$$\int_a^b \int_{c(x)}^{d(x)} f(x, y)\, dy\, dx \tag{4.42}$$

o

$$\int_c^d \int_{a(y)}^{b(y)} f(x, y)\, dx\, dy. \tag{4.43}$$

De hecho, las integrales en las regiones que no son de este tipo, también se pueden aproximar al realizar subdivisiones adecuadas de la región. (Consulte el ejercicio 10.)

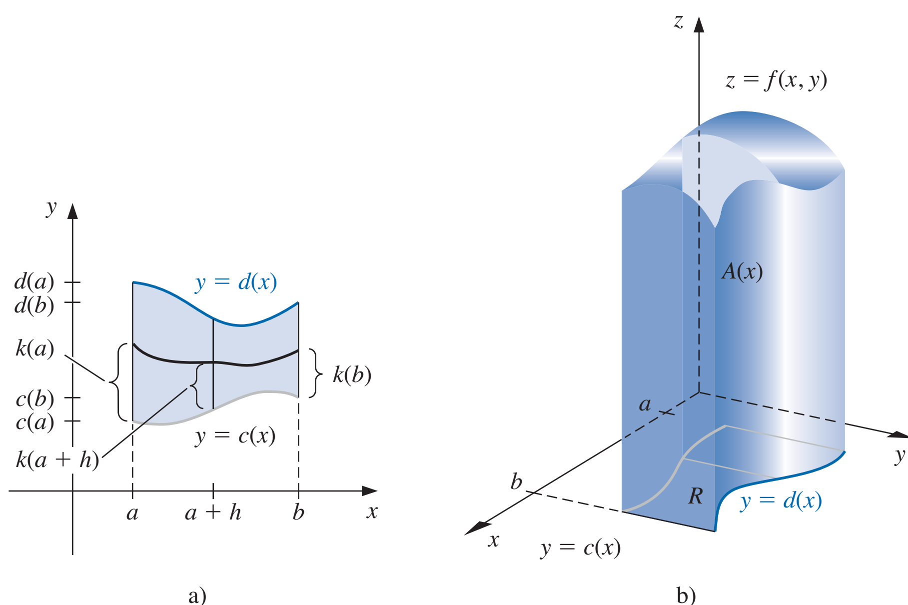
Para describir la técnica relacionada con la aproximación de una integral de la forma

$$\int_a^b \int_{c(x)}^{d(x)} f(x, y)\, dy\, dx,$$

utilizaremos la regla básica de Simpson para integrar respecto a ambas variables. El tamaño de paso para la variable x es $h = (b - a)/2$, pero el tamaño de paso para y varía con x (véase la figura 4.21) y se escribe

$$k(x) = \frac{d(x) - c(x)}{2}.$$

Figura 4.21



Esto nos da

$$\int_a^b \int_{c(x)}^{d(x)} f(x, y)\,dy\,dx \approx \int_a^b \frac{k(x)}{3}[f(x, c(x)) + 4f(x, c(x) + k(x)) + f(x, d(x))]\,dx$$

$$\approx \frac{h}{3}\left\{ \frac{k(a)}{3}[f(a, c(a)) + 4f(a, c(a) + k(a)) + f(a, d(a))] \right.$$

$$+ \frac{4k(a+h)}{3}[f(a+h, c(a+h)) + 4f(a+h, c(a+h)$$

$$+ k(a+h)) + f(a+h, d(a+h))]$$

$$\left. + \frac{k(b)}{3}[f(b, c(b)) + 4f(b, c(b) + k(b)) + f(b, d(b))] \right\}.$$

El algoritmo 4.4. aplica la regla compuesta de Simpson para una integral de la forma (4.42). Las integrales de la forma (4.43) pueden, por supuesto, manejarse de manera similar.

ALGORITMO **4.4**

Integral doble de Simpson

Para aproximar la integral

$$I = \int_a^b \int_{c(x)}^{d(x)} f(x, y)\,dy\,dx:$$

ENTRADA extremos a, b: enteros positivos pares m, n.

SALIDA aproximación J para I.

Paso 1 Haga $h = (b-a)/n$;
$J_1 = 0$; (*Términos finales.*)
$J_2 = 0$; (*Términos pares.*)
$J_3 = 0$. (*Términos impares.*)

Paso 2 Para $i = 0, 1, \dots, n$ haga los pasos 3–8.

Paso 3 Haga $x = a + ih$; (*Método compuesto de Simpson para x.*)
$HX = (d(x) - c(x))/m$;
$K_1 = f(x, c(x)) + f(x, d(x))$; (*Términos finales.*)
$K_2 = 0$; (*Términos pares.*)
$K_3 = 0$. (*Términos impares.*)

Paso 4 Para $j = 1, 2, \dots, m-1$ haga los pasos 5 y 6.

Paso 5 Haga $y = c(x) + jHX$;
$Q = f(x, y)$.

Paso 6 Si j incluso determine entonces $K_2 = K_2 + Q$
también determine $K_3 = K_3 + Q$.

Paso 7 Haga $L = (K_1 + 2K_2 + 4K_3)HX/3$.

$$\left(L \approx \int_{c(x_i)}^{d(x_i)} f(x_i, y)\,dy \quad \textit{mediante el método compuesto de Simpson}. \right)$$

Paso 8 Si $i = 0$ o $i = n$ entonces detemine $J_1 = J_1 + L$
también si i incluso determine entonces $J_2 = J_2 + L$
también determine $J_3 = J_3 + L$. (*Paso final 2*)

Paso 9 Haga $J = h(J_1 + 2J_2 + 4J_3)/3$.

Paso 10 SALIDA (J);
PARE. ■

Aplicando la cuadratura gaussiana a la integral doble

$$\int_a^b \int_{c(x)}^{d(x)} f(x, y)\, dy\, dx$$

primero debemos transformar, para cada x en $[a, b]$, la variable y en el intervalo $[c(x), d(x)]$ en la variable t en el intervalo $[-1, 1]$. Esta transformación lineal nos da

$$f(x, y) = f\left(x, \frac{(d(x) - c(x))t + d(x) + c(x)}{2}\right) \quad \text{y} \quad dy = \frac{d(x) - c(x)}{2}\, dt.$$

El cálculo reducido hace que, en general, valga la pena aplicar la cuadratura gaussiana en lugar de una técnica de Simpson al aproximar las integrales dobles.

Entonces, para cada x en $[a, b]$, aplicamos la cuadratura gaussiana a la integral resultante

$$\int_{c(x)}^{d(x)} f(x, y)\, dy = \int_{-1}^{1} f\left(x, \frac{(d(x) - c(x))t + d(x) + c(x)}{2}\right) dt$$

para producir

$$\int_a^b \int_{c(x)}^{d(x)} f(x, y)\, dy\, dx$$
$$\approx \int_a^b \frac{d(x) - c(x)}{2} \sum_{j=1}^{n} c_{n,j} f\left(x, \frac{(d(x) - c(x))r_{n,j} + d(x) + c(x)}{2}\right) dx,$$

mientras, como antes, las raíces $r_{n,j}$ y los coeficientes $c_{n,j}$ provienen de la tabla 4.12 en la página 172. Ahora, el intervalo $[a, b]$ se transforma en $[-1, 1]$, y la cuadratura gaussiana se aplica para aproximar la integral en el lado derecho de esta ecuación. Los detalles se incluyen en el algoritmo 4.5.

ALGORITMO **4.5**

Integral doble gaussiana

Para aproximar la integral

$$\int_a^b \int_{c(x)}^{d(x)} f(x, y)\, dy\, dx:$$

ENTRADA extremos a, b: enteros positivos m, n.
(*Las raíces $r_{i,j}$ y los coeficientes $c_{i,j}$ necesitan estar disponibles para* $i = \text{máx}\{m, n\}$ *y para* $1 \leq j \leq i$.)

SALIDA aproximación J para I.

Paso 1 Haga $h_1 = (b - a)/2$;
$h_2 = (b + a)/2$;
$J = 0$.

Paso 2 Para $i = 1, 2, \ldots, m$ haga los pasos 3–5.

Paso 3 Haga $JX = 0$;
$x = h_1 r_{m,i} + h_2$;
$d_1 = d(x)$;
$c_1 = c(x)$;
$k_1 = (d_1 - c_1)/2$;
$k_2 = (d_1 + c_1)/2$.

Paso 4 Para $j = 1, 2, \ldots, n$ haga
$y = k_1 r_{n,j} + k_2$;
$Q = f(x, y)$;
$JX = JX + c_{n,j} Q$.

Paso 5 Haga $J = J + c_{m,i} k_1 J X$. *(Paso final 2.)*

Paso 6 Haga $J = h_1 J$.

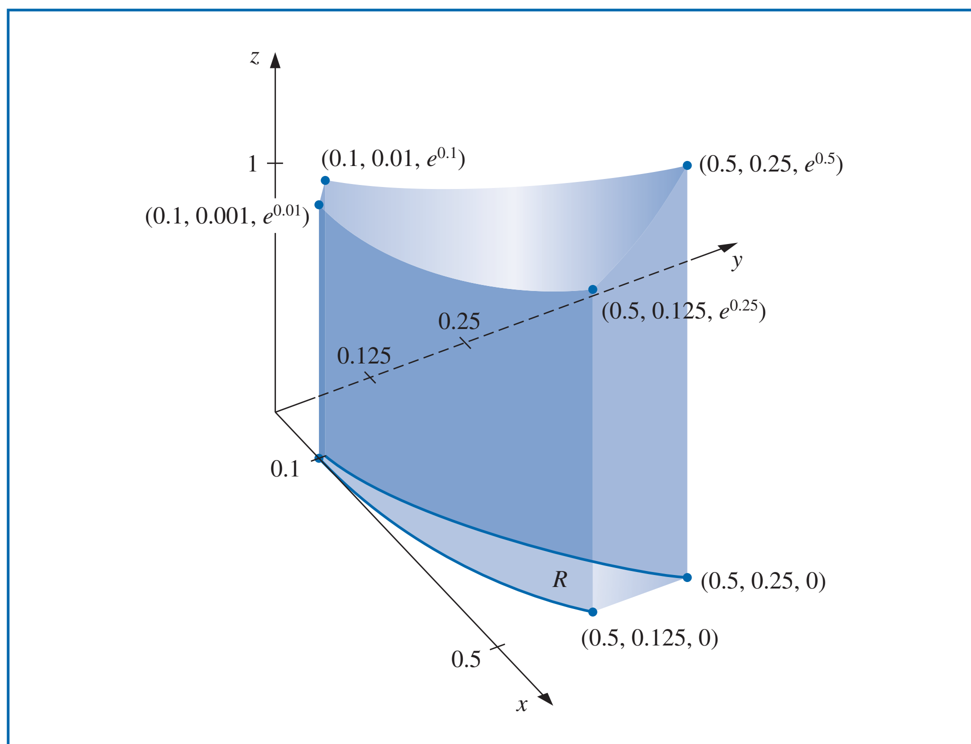
Paso 7 SALIDA (J);
PARE. ■

Ilustración El volumen del sólido en la figura 4.22 se aproxima aplicando el algoritmo de la integral doble de Simpson con $n = m = 10$ a

$$\int_{0.1}^{0.5} \int_{x^3}^{x^2} e^{y/x}\, dy\, dx.$$

Esto requiere 121 evaluaciones de la función $f(x, y) = e^{y/x}$ y produce el valor 0.0333054, el cual aproxima el volumen del sólido mostrado en la figura 4.22 a aproximadamente siete lugares decimales. Al aplicar el algoritmo de cuadratura gaussiana con $n = m = 5$ sólo se requieren 25 evaluaciones de función y da la aproximación 0.03330556611, lo cual es preciso para 11 lugares decimales. ■

Figura 4.22



Aproximación de integral triple

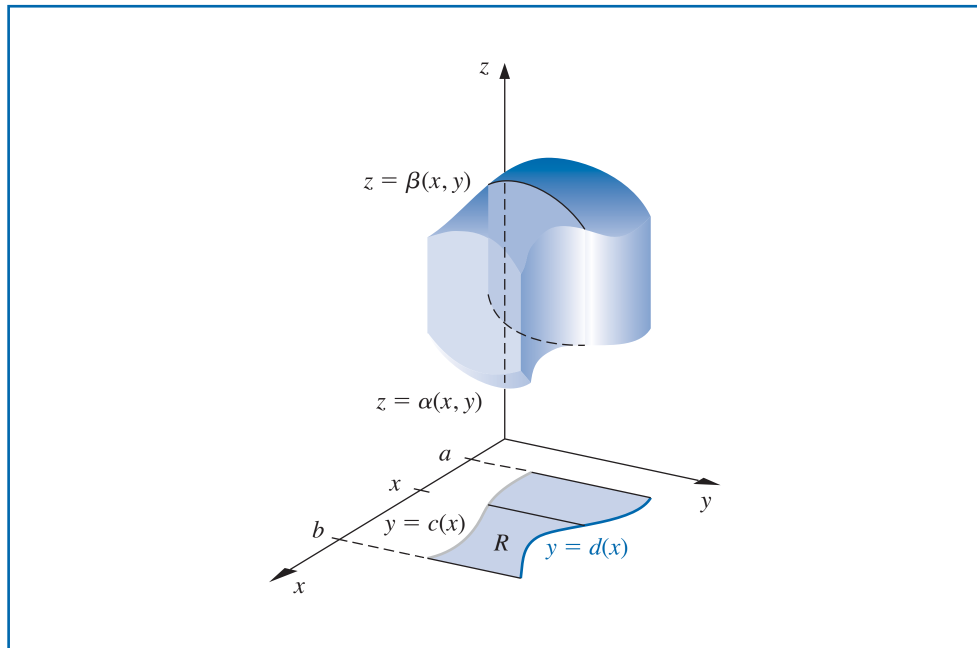
Las integrales triples de la forma

$$\int_{a}^{b} \int_{c(x)}^{d(x)} \int_{\alpha(x,y)}^{\beta(x,y)} f(x, y, z)\, dz\, dy\, dx$$

El cálculo reducido casi siempre hace que valga la pena aplicar la cuadratura gaussiana en lugar de la técnica de Simpson al aproximar integrales triples o superiores.

(Véase la figura 4.23) se aproximan de manera similar. Debido al número de cálculos implicados, la cuadratura gaussiana es el método de elección. El algoritmo 4.6 implementa este procedimiento.

Figura 4.23



ALGORITMO 4.6

Integral triple gaussiana

Para aproximar la integral

$$\int_a^b \int_{c(x)}^{d(x)} \int_{\alpha(x,y)}^{\beta(x,y)} f(x, y, z)\, dz\, dy\, dx :$$

ENTRADA extremos a, b; enteros positivos m, n, p.
(*Las raíces* $r_{i,j}$ *y los coeficientes* $c_{i,j}$ *necesitan estar disponibles para* $i = \text{máx}\{n, m, p\}$ *y para* $1 \leq j \leq i$.)

SALIDA aproximación J para I.

Paso 1 Haga $h_1 = (b - a)/2$;
$h_2 = (b + a)/2$;
$J = 0$.

Paso 2 Para $i = 1, 2, \ldots, m$ haga los pasos 3–8.

Paso 3 Haga $JX = 0$;
$x = h_1 r_{m,i} + h_2$;
$d_1 = d(x)$;
$c_1 = c(x)$;
$k_1 = (d_1 - c_1)/2$;
$k_2 = (d_1 + c_1)/2$.

Paso 4 Para $j = 1, 2, \ldots, n$ haga los pasos 5–7.

Paso 5 Haga $JY = 0$;
$y = k_1 r_{n,j} + k_2$;
$\beta_1 = \beta(x, y)$;
$\alpha_1 = \alpha(x, y)$;
$l_1 = (\beta_1 - \alpha_1)/2$;
$l_2 = (\beta_1 + \alpha_1)/2$.

Paso 6 Para $k = 1, 2, \ldots, p$ haga
Haga $z = l_1 r_{p,k} + l_2$;
$Q = f(x, y, z)$;
$JY = JY + c_{p,k} Q$.

Paso 7 Haga $JX = JX + c_{n,j} l_1 JY$. *(Fin del paso 4)*

Paso 8 Haga $J = J + c_{m,i} k_1 JX$. *(Fin del paso 2)*

Paso 9 Haga $J = h_1 J$.

Paso 10 SALIDA (J);
PARE.

El siguiente ejemplo requiere la evaluación de cuatro integrales triples.

Ilustración El centro de masa de una región sólida D con función de densidad σ se presenta en

$$(\overline{x}, \overline{y}, \overline{z}) = \left(\frac{M_{yz}}{M}, \frac{M_{xz}}{M}, \frac{M_{xy}}{M} \right),$$

donde

$$M_{yz} = \iiint_D x\sigma(x, y, z)\, dV, \quad M_{xz} = \iiint_D y\sigma(x, y, z)\, dV$$

y

$$M_{xy} = \iiint_D z\sigma(x, y, z)\, dV$$

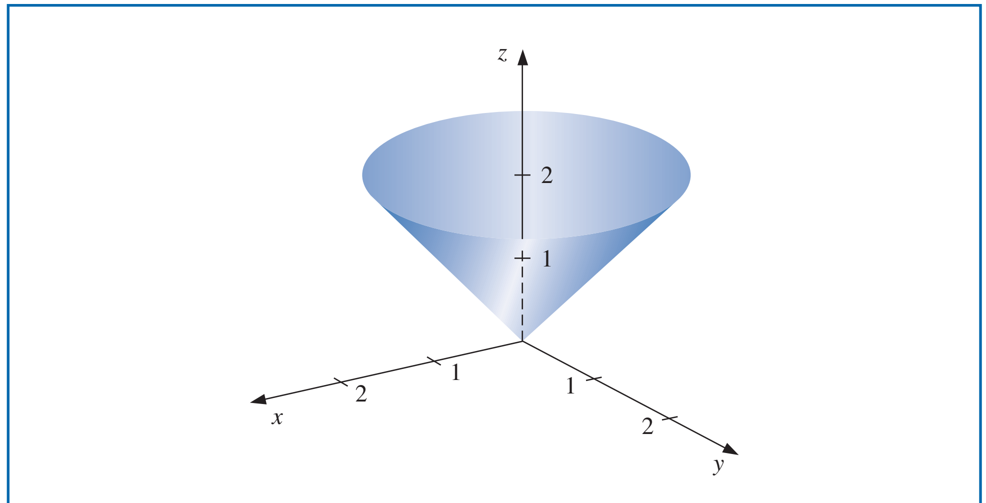
son los momentos alrededor de los planos coordenados y la masa de D es

$$M = \iiint_D \sigma(x, y, z)\, dV.$$

El sólido mostrado en la figura 4.24 está acotado por la parte superior del cono $z^2 = x^2 + y^2$ y el plano $z = 2$. Suponga que este sólido tiene una función de densidad dada por

$$\sigma(x, y, z) = \sqrt{x^2 + y^2}.$$

Figura 4.24



Al aplicar el algoritmo de integral triple gaussiana 4.6 con $n = m = p = 5$ requiere 125 evaluaciones de función por integral y se obtienen las siguientes aproximaciones

$$M = \int_{-2}^{2} \int_{-\sqrt{4-x^2}}^{\sqrt{4-x^2}} \int_{\sqrt{x^2+y^2}}^{2} \sqrt{x^2+y^2}\, dz\, dy\, dx$$

$$= 4 \int_{0}^{2} \int_{0}^{\sqrt{4-x^2}} \int_{\sqrt{x^2+y^2}}^{2} \sqrt{x^2+y^2}\, dz\, dy\, dx \approx 8.37504476,$$

$$M_{yz} = \int_{-2}^{2} \int_{-\sqrt{4-x^2}}^{\sqrt{4-x^2}} \int_{\sqrt{x^2+y^2}}^{2} x\sqrt{x^2+y^2}\, dz\, dy\, dx \approx -5.55111512 \times 10^{-17},$$

$$M_{xz} = \int_{-2}^{2} \int_{-\sqrt{4-x^2}}^{\sqrt{4-x^2}} \int_{\sqrt{x^2+y^2}}^{2} y\sqrt{x^2+y^2}\, dz\, dy\, dx \approx -8.01513675 \times 10^{-17} \quad \text{y}$$

$$M_{xy} = \int_{-2}^{2} \int_{-\sqrt{4-x^2}}^{\sqrt{4-x^2}} \int_{\sqrt{x^2+y^2}}^{2} z\sqrt{x^2+y^2}\, dz\, dy\, dx \approx 13.40038156.$$

Esto implica que la ubicación aproximada del centro de la masa es

$$(\overline{x}, \overline{y}, \overline{z}) = (0, 0, 1.60003701).$$

Estas integrales son bastante fáciles de evaluar de manera directa. Si lo hace, descubrirá que el centro de masa exacto se presenta en (0, 0, 1.6). ■

La sección Conjunto de ejercicios 4.8 está disponible en línea. Encuentre la ruta de acceso en las páginas preliminares.

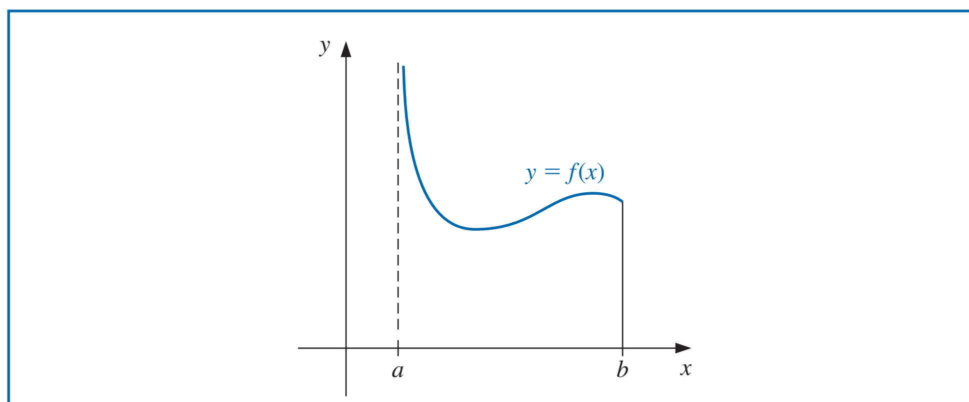
4.9 Integrales impropias

Las integrales impropias resultan cuando la noción de integración se amplía, ya sea en un intervalo de integración en el que la función no está acotada en un intervalo con uno o más extremos infinitos. En cualquier circunstancia, las reglas normales de una aproximación integral se deben modificar.

Singularidad del extremo izquierdo

Primero consideraremos la situación en la que el integrando no está limitado en el extremo izquierdo del intervalo de integración, como se muestra en la figura 4.25. En este caso, decimos que f tiene una **singularidad** en el extremo a. A continuación, mostramos cómo se pueden reducir las integrales impropias para problemas de este tipo.

Figura 4.25



En cálculo se muestra que la integral impropia con una singularidad en el extremo izquierdo,

$$\int_a^b \frac{dx}{(x-a)^p},$$

converge si y sólo si $0 < p < 1$, y en este caso, definimos

$$\int_a^b \frac{1}{(x-a)^p}\,dx = \lim_{M\to a^+} \left.\frac{(x-a)^{1-p}}{1-p}\right|_{x=M}^{x=b} = \frac{(b-a)^{1-p}}{1-p}.$$

Ejemplo 1 Muestre que la integral impropia $\int_0^1 \frac{1}{\sqrt{x}}\,dx$ converge pero que $\int_0^1 \frac{1}{x^2}\,dx$ diverge.

Solución Para la primera integral, tenemos

$$\int_0^1 \frac{1}{\sqrt{x}}\,dx = \lim_{M\to 0^+}\int_M^1 x^{-1/2}\,dx = \lim_{M\to 0^+} 2x^{1/2}\Big|_{x=M}^{x=1} = 2-0=2,$$

pero la segunda integral

$$\int_0^1 \frac{1}{x^2}\,dx = \lim_{M\to 0^+}\int_M^1 x^{-2}\,dx = \lim_{M\to 0^+} -x^{-1}\Big|_{x=M}^{x=1}$$

es infinita. ■

Si f es una función que se puede escribir en la forma

$$f(x) = \frac{g(x)}{(x-a)^p},$$

donde $0 < p < 1$ y g es continua en $[a, b]$, entonces la integral impropia

$$\int_a^b f(x)\,dx$$

también existe. Nosotros aproximaremos esta integral mediante la regla compuesta de Simpson, siempre que $g \in C^5[a,b]$. En ese caso, construimos el cuarto polinomio de Taylor $P_4(x)$, para g alrededor de a,

$$P_4(x) = g(a) + g'(a)(x-a) + \frac{g''(a)}{2!}(x-a)^2 + \frac{g'''(a)}{3!}(x-a)^3 + \frac{g^{(4)}(a)}{4!}(x-a)^4,$$

y escribimos

$$\int_a^b f(x)\,dx = \int_a^b \frac{g(x)-P_4(x)}{(x-a)^p}\,dx + \int_a^b \frac{P_4(x)}{(x-a)^p}\,dx. \tag{4.44}$$

Puesto que $P(x)$ es un polinomio, podemos determinar con exactitud el valor de

$$\int_a^b \frac{P_4(x)}{(x-a)^p}\,dx = \sum_{k=0}^{4}\int_a^b \frac{g^{(k)}(a)}{k!}(x-a)^{k-p}\,dx = \sum_{k=0}^{4} \frac{g^{(k)}(a)}{k!(k+1-p)}(b-a)^{k+1-p}. \tag{4.45}$$

Por lo general, la parte dominante de la aproximación, en especial cuando el polinomio de Taylor $P_4(x)$ concuerda de cerca con $g(x)$ en todo el intervalo $[a, b]$.

Para aproximar la integral de f, debemos añadir este valor a la aproximación de

$$\int_a^b \frac{g(x) - P_4(x)}{(x-a)^p}\, dx.$$

Para determinar esto, primero definimos

$$G(x) = \begin{cases} \frac{g(x)-P_4(x)}{(x-a)^p}, & \text{si } a < x \leq b, \\ 0, & \text{si } x = a. \end{cases}$$

Esto nos da una función continua en $[a, b]$. De hecho, $0 < p < 1$ y $P_4^{(k)}(a)$ concuerda con $g^{(k)}(a)$ para cada $k = 0, 1, 2, 3, 4,$ por lo que tenemos $G \in C^4[a, b]$. Esto implica que la regla compuesta de Simpson se puede aplicar para aproximar la integral de G sobre $[a, b]$. Al añadir esta aproximación al valor de la ecuación (4.45) obtenemos una aproximación para la integral impropia de f en $[a, b]$, dentro de la precisión de la aproximación de la regla compuesta de Simpson.

Ejemplo 2 Utilice la regla compuesta de Simpson con $h = 0.25$ para aproximar el valor de la integral impropia

$$\int_0^1 \frac{e^x}{\sqrt{x}}\, dx.$$

Solución El cuarto polinomio de Taylor para e^x alrededor de $x = 0$ es

$$P_4(x) = 1 + x + \frac{x^2}{2} + \frac{x^3}{6} + \frac{x^4}{24},$$

por lo que la parte dominante de la aproximación para $\int_0^1 \frac{e^x}{\sqrt{x}}\, dx$ es

$$\int_0^1 \frac{P_4(x)}{\sqrt{x}}\, dx = \int_0^1 \left(x^{-1/2} + x^{1/2} + \frac{1}{2}x^{3/2} + \frac{1}{6}x^{5/2} + \frac{1}{24}x^{7/2} \right) dx$$

$$= \lim_{M\to 0^+} \left[2x^{1/2} + \frac{2}{3}x^{3/2} + \frac{1}{5}x^{5/2} + \frac{1}{21}x^{7/2} + \frac{1}{108}x^{9/2} \right]_M^1$$

$$= 2 + \frac{2}{3} + \frac{1}{5} + \frac{1}{21} + \frac{1}{108} \approx 2.9235450.$$

Para la segunda parte de la aproximación para $\int_0^1 \frac{e^x}{\sqrt{x}}\, dx$, necesitamos aproximar $\int_0^1 G(x)\, dx$, donde

$$G(x) = \begin{cases} \frac{1}{\sqrt{x}}\left(e^x - P_4(x)\right), & \text{si } 0 < x \leq 1, \\ 0, & \text{si } x = 0. \end{cases}$$

Tabla 4.13

x	$G(x)$
0.00	0
0.25	0.0000170
0.50	0.0004013
0.75	0.0026026
1.00	0.0099485

La tabla 4.13 enumera los valores necesarios para la regla compuesta de Simpson para esta aproximación.

Por medio de estos datos y la regla compuesta de Simpson obtenemos

$$\int_0^1 G(x)\, dx \approx \frac{0.25}{3}[0 + 4(0.0000170) + 2(0.0004013) + 4(0.0026026) + 0.0099485]$$

$$= 0.0017691.$$

Por lo tanto,

$$\int_0^1 \frac{e^x}{\sqrt{x}}\, dx \approx 2.9235450 + 0.0017691 = 2.9253141.$$

Este resultado es preciso dentro de la precisión de la aproximación de la regla compuesta de Simpson para la función G. Puesto que $|G^{(4)}(x)| < 1$ en $[0, 1]$, el error está acotado por

$$\frac{1-0}{180}(0.25)^4 = 0.0000217.$$ ■

Singularidad en el extremo derecho

Para aproximar la integral impropia con una singularidad en el extremo derecho, podemos desarrollar una técnica similar, pero expandiendo los términos del extremo derecho b en lugar del extremo izquierdo a. De forma alternativa, podemos realizar la sustitución

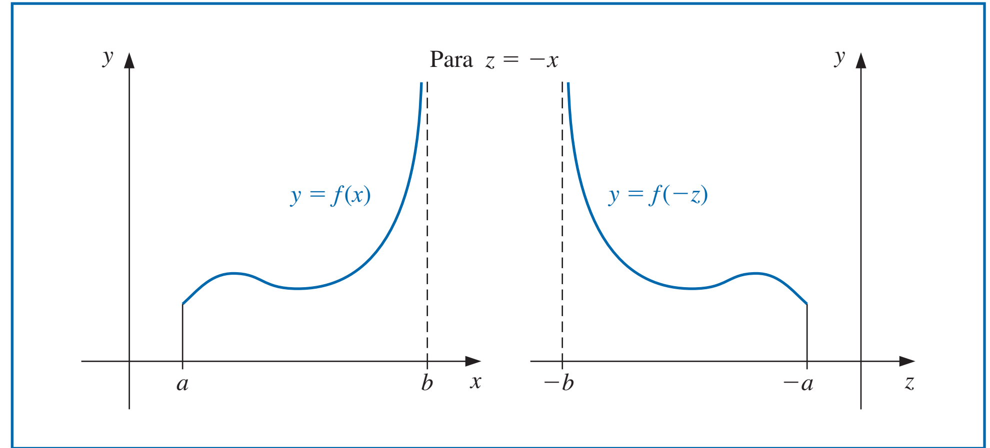
$$z = -x, \qquad dz = -\, dx$$

para cambiar la integral impropia en una de la forma

$$\int_a^b f(x)\, dx = \int_{-b}^{-a} f(-z)\, dz, \tag{4.46}$$

la cual tiene su singularidad en el extremo izquierdo. Entonces, podemos aplicar la técnica de singularidad del extremo izquierdo que ya habíamos desarrollado. (Véase la figura 4.26.)

Figura 4.26



Una integral impropia con una singularidad en c, donde $a < c < b$, se trata como la suma de las integrales impropias con singularidades en los extremos ya que

$$\int_a^b f(x)\, dx = \int_a^c f(x)\, dx + \int_c^b f(x)\, dx.$$

Singularidad infinita

El otro tipo de integral impropia implica límites infinitos de integración. La integral básica de este tipo tiene la forma

$$\int_a^\infty \frac{1}{x^p}\, dx,$$

para $p > 1$. Esto se convierte en una integral con singularidad de extremo izquierdo en 0 al realizar la sustitución de integración

$$t = x^{-1}, \quad dt = -x^{-2}\,dx, \quad \text{por lo que} \quad dx = -x^2\,dt = -t^{-2}\,dt.$$

Entonces

$$\int_a^\infty \frac{1}{x^p}\,dx = \int_{1/a}^0 -\frac{t^p}{t^2}\,dt = \int_0^{1/a} \frac{1}{t^{2-p}}\,dt.$$

De forma similar, el cambio de variable $t = x^{-1}$ convierte la integral impropia $\int_a^\infty f(x)\,dx$ en una que tiene singularidad de extremo izquierdo en cero:

$$\int_a^\infty f(x)\,dx = \int_0^{1/a} t^{-2} f\left(\frac{1}{t}\right)\,dt. \tag{4.47}$$

Ahora, se puede aproximar por medio de la fórmula de cuadratura del tipo descrito anteriormente.

Ejemplo 3 Aproxime el valor de la integral impropia

$$I = \int_1^\infty x^{-3/2} \operatorname{sen} \frac{1}{x}\,dx.$$

Solución Primero realizamos el cambio de variable $t = x^{-1}$, el cual convierte la singularidad infinita en una con una singularidad de extremo izquierdo. Entonces

$$dt = -x^{-2}\,dx, \quad \text{por lo que} \quad dx = -x^2\,dt = -\frac{1}{t^2}\,dt,$$

y

$$I = \int_{x=1}^{x=\infty} x^{-3/2} \operatorname{sen} \frac{1}{x}\,dx = \int_{t=1}^{t=0} \left(\frac{1}{t}\right)^{-3/2} \operatorname{sen} t \left(-\frac{1}{t^2}\,dt\right) = \int_0^1 t^{-1/2} \operatorname{sen} t\,dt.$$

El cuarto polinomio de Taylor, $P_4(t)$, para *sen t* alrededor de 0 es

$$P_4(t) = t - \frac{1}{6}t^3,$$

por lo tanto,

$$G(t) = \begin{cases} \dfrac{\operatorname{sen} t - t + \frac{1}{6}t^3}{t^{1/2}}, & \text{si } 0 < t \le 1 \\ 0, & \text{si } t = 0 \end{cases}$$

está en $C^4[0, 1]$, y tenemos

$$\begin{aligned} I &= \int_0^1 t^{-1/2}\left(t - \frac{1}{6}t^3\right)\,dt + \int_0^1 \frac{\operatorname{sen} t - t + \frac{1}{6}t^3}{t^{1/2}}\,dt \\ &= \left[\frac{2}{3}t^{3/2} - \frac{1}{21}t^{7/2}\right]_0^1 + \int_0^1 \frac{\operatorname{sen} t - t + \frac{1}{6}t^3}{t^{1/2}}\,dt \\ &= 0.61904761 + \int_0^1 \frac{\operatorname{sen} t - t + \frac{1}{6}t^3}{t^{1/2}}\,dt. \end{aligned}$$

El resultado a partir de la regla compuesta de Simpson con $n = 16$ para la integral restante es 0.0014890097. Esto da una aproximación final de

$$I = 0.0014890097 + 0.61904761 = 0.62053661,$$

lo cual es preciso dentro de 4.0×10^{-8}. ■

La sección Conjunto de ejercicios 4.9 está disponible en línea. Encuentre la ruta de acceso en las páginas preliminares.

4.10 Software numérico y revisión del capítulo

La mayoría del software para integrar una función de una única variable real está basado ya sea en el enfoque adaptable o en fórmulas gaussianas extremadamente precisas. La integración cautelosa de Romberg es una técnica adaptable que incluye una verificación para garantizar que el integrando se comporta suavemente sobre los subintervalos de la integral de integración. Este método se ha utilizado con éxito en bibliotecas de software. En general, múltiples integrales se aproximan con la ampliación de buenos métodos adaptables hasta dimensiones superiores. La cuadratura del tipo gaussiano también se recomienda para disminuir el número de evaluaciones de función.

Las rutinas principales en las bibliotecas IMSL y NAG están basadas en QUADPACK: A Subroutine Package for Automatic Integration (QUADPACK: *Un paquete de subrutinas para integración automática*) de R. Piessens, E. de Doncker-Kapenga, C. W. Uberhuber y D. K. Kahaner y publicado por Springer-Verlag en 1983 [PDUK].

La biblioteca IMSL contiene un esquema de integración adaptable con base en la regla Gaussiana-Kronrod de 21 puntos mediante la regla gaussiana de 10 puntos para cálculo de error. Las reglas gaussianas utilizan 10 puntos $x_1, \ldots, x_{10}$ y pesos $w_1, \ldots, w_{10}$ para la fórmula de cuadratura $\sum_{i=1}^{10} w_i f(x_i)$ para aproximar $\int_a^b f(x)\,dx$. A continuación se utilizan los puntos adicionales $x_{11}, \ldots, x_{21}$, y los pesos nuevos $v_1, \ldots, v_{21}$, en la fórmula de Kronrod $\sum_{i=1}^{21} v_i f(x_i)$. Los resultados de las dos fórmulas se comparan para eliminar el error. La ventaja de utilizar $x_1, \ldots, x_{10}$ en cada fórmula es que f sólo necesita evaluarse en 21 puntos. Si se utilizaran reglas gaussianas independientes de 10 y 21 puntos, se necesitarían 31 evaluaciones de función. Este procedimiento permite singularidades de extremo en el integrando.

Otras subrutinas IMSL permiten singularidades de extremo, singularidades especificadas por el usuario e intervalos infinitos de integración. Además, existen rutinas para aplicar reglas de Gauss-Kronrod para integrar una función de dos variables y una rutina para utilizar cuadratura para integrar una función de n variables sobre n intervalos de la forma $[a_i, b_i]$.

La Biblioteca NAG incluye una rutina para calcular la integral de f sobre el intervalo $[a, b]$ mediante un método adaptable con base en cuadratura gaussiana mediante reglas de Kronrod de 21 puntos y de Gauss de 10 puntos. También tiene una rutina para aproximar una integral mediante una familia de fórmulas tipo gaussianas con base en 1, 3, 5, 7, 15, 31, 63, 127 y 255 nodos. Estas reglas entrelazadas de alta precisión se deben a Patterson [Pat] y se utilizan de manera adaptable. NAG incluye muchas otras subrutinas para aproximar integrales.

A pesar de que la diferenciación numérica es inestable, se necesitan fórmulas de aproximación de derivadas para resolver ecuaciones diferenciales. La Biblioteca NAG incluye una subrutina para la diferenciación numérica de una función de una variable real con diferenciación para que la catorceava derivada sea posible. IMSL tiene una función que usa un cambio adaptable en tamaño de paso para diferencias finitas para aproximar la primera, segunda o tercera derivada de f en x dentro de una tolerancia determinada. IMSL también incluye una subrutina para calcular las derivadas de una función definida en un conjunto de puntos me-

diante interpolación cuadrática. Ambos paquetes permiten la diferenciación e integración de splines cúbicos interpolantes construidos por las subrutinas mencionadas en la sección 3.5.

Las secciones Preguntas de análisis, Conceptos clave y Revisión del capítulo están disponibles en línea. Encuentre la ruta de acceso en las páginas preliminares.

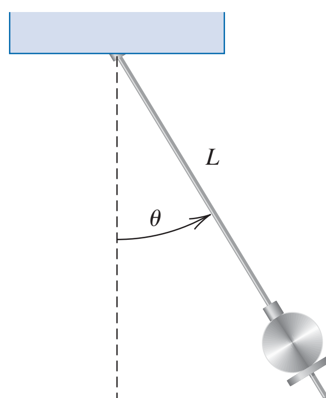
CAPÍTULO

5 Problemas de valor inicial para ecuaciones de diferenciales ordinarias

Introducción

El movimiento de un péndulo balanceándose de acuerdo con ciertas suposiciones de simplificación se describe por medio de la ecuación diferencial de segundo orden.

$$\frac{d^2\theta}{dt^2} + \frac{g}{L}\operatorname{sen}\theta = 0,$$



donde L es la longitud del péndulo, $g \approx 32.17$ pies/s^2 es la constante gravitacional de la Tierra, y θ es el ángulo del péndulo con la vertical. Si, además, especificamos la posición del péndulo cuando el movimiento empieza, $\theta(t_0) = \theta_0$, y su velocidad en ese punto, $\theta'(t_0) = \theta'_0$, tenemos lo que recibe el nombre de *problema de valor inicial.*

Para los valores pequeños de θ, la aproximación $\theta \approx \operatorname{sen}\theta$ se puede utilizar para simplificar el problema de valor inicial lineal

$$\frac{d^2\theta}{dt^2} + \frac{g}{L}\theta = 0, \quad \theta(t_0) = \theta_0, \quad \theta'(t_0) = \theta'_0.$$

Este problema se puede resolver con una técnica de ecuación diferencial estándar. Para los valores más grandes de θ, la suposición de que $\theta = \operatorname{sen}\theta$ no es razonable, por lo que deben usarse los métodos de aproximación. Un problema de este tipo se considera en el ejercicio 7 de la sección 5.9.

Cualquier libro de texto sobre ecuaciones diferenciales detalla diferentes métodos para encontrar soluciones a los problemas de valor inicial de primer orden de manera explícita. Sin embargo, en la práctica, pocos problemas que se originan a partir del estudio de fenómenos físicos se pueden resolver con exactitud.

en su segunda variable, y la condición (5.1) es generalmente más fácil de aplicar que la definición. Sin embargo, deberíamos observar que el teorema 5.3 solamente proporciona condiciones suficientes para mantener la condición de Lipschitz. La función en el ejemplo 1, por ejemplo, satisface la condición de Lipschitz, pero la derivada parcial respecto a y no existe cuando $y = 0$.

El siguiente teorema es una versión del teorema fundamental de existencia y unicidad para ecuaciones diferenciales ordinarias de primer orden. A pesar de que el teorema se puede probar reduciendo la hipótesis de alguna forma, esta versión es suficiente para nuestros propósitos. (La prueba del teorema, enunciado de esta forma, se puede encontrar en [BiR], pp. 142–155.)

Teorema 5.4 Suponga que $D = \{(t, y) \mid a \leq t \leq b \text{ y } -\infty < y < \infty\}$ y que $f(t, y)$ es continua en D. Si f satisface la condición de Lipschitz en D en la variable y, entonces el problema de valor inicial

$$y'(t) = f(t, y), \quad a \leq t \leq b, \quad y(a) = \alpha,$$

tiene una única solución $y(t)$ para $a \leq t \leq b$. ■

Ejemplo 2 Utilice el teorema 5.4 para mostrar que existe una única solución para el problema de valor inicial

$$y' = 1 + t\,\text{sen}(ty), \quad 0 \leq t \leq 2, \quad y(0) = 0.$$

Solución Al mantener t constante y aplicar el teorema de valor medio para la función

$$f(t, y) = 1 + t\,\text{sen}(ty),$$

encontramos que cuando $y_1 < y_2$, existe un número ξ en (y_1, y_2) con

$$\frac{f(t, y_2) - f(t, y_1)}{y_2 - y_1} = \frac{\partial}{\partial y} f(t, \xi) = t^2 \cos(\xi t).$$

Por lo tanto,

$$|f(t, y_2) - f(t, y_1)| = |y_2 - y_1||t^2 \cos(\xi t)| \leq 4|y_2 - y_1|,$$

y f satisface la condición de Lipschitz en la variable y con constante de Lipschitz $L = 4$. Además, $f(t, y)$ es continua cuando $0 \leq t \leq 2$ y $-\infty < y < \infty$, por lo que el teorema 5.4 implica que existe una única solución para este problema de valor inicial.

Si usted ha completado un curso sobre ecuaciones diferenciales, podría intentar encontrar la solución exacta para este problema. ■

Problemas bien planteados

Ahora que hemos, hasta cierto punto, atendido la cuestión de cuando los problemas de valor inicial tienen soluciones únicas, podemos mover la segunda consideración importante al aproximar la solución para un problema de valor inicial. En general, los problemas de valor inicial obtenidos al observar el fenómeno físico solamente aproximan la verdadera situación, por lo que necesitamos saber si los pequeños cambios en la declaración del problema introducen pequeños cambios en la solución en la misma medida. Esto también es importante debido a la introducción del error de redondeo cuando se usan métodos numéricos. Esto es,

- Pregunta: ¿Cómo determinamos si un problema particular tiene la propiedad de que pequeños cambios o alteraciones del problema introduzcan pequeños cambios en la solución en la misma medida?

Como siempre, primero necesitamos proporcionar una definición práctica para expresar este concepto.

La primera parte de este capítulo se preocupa por aproximar la solución $y(t)$ para un problema de la forma

$$\frac{dy}{dt} = f(t, y), \quad \text{para } a \leq t \leq b,$$

sujeto a la condición inicial $y(a) = \alpha$. Más adelante en este capítulo tratamos con la extensión de estos métodos hacia un sistema de ecuaciones diferenciales de primer orden de la forma

$$\frac{dy_1}{dt} = f_1(t, y_1, y_2, \ldots, y_n),$$
$$\frac{dy_2}{dt} = f_2(t, y_1, y_2, \ldots, y_n),$$
$$\vdots$$
$$\frac{dy_n}{dt} = f_n(t, y_1, y_2, \ldots, y_n),$$

para $a \leq t \leq b$, sujeto a las condiciones iniciales

$$y_1(a) = \alpha_1, \quad y_2(a) = \alpha_2, \quad \ldots, \quad y_n(a) = \alpha_n.$$

También examinamos la relación de un sistema de este tipo con el problema de valor inicial de enésimo orden de la forma

$$y^{(n)} = f(t, y, y', y'', \ldots, y^{(n-1)}),$$

para $a \leq t \leq b$, sujeto a las condiciones iniciales

$$y(a) = \alpha_1, \quad y'(a) = \alpha_2, \quad \ldots, \quad y^{n-1}(a) = \alpha_n.$$

5.1 Teoría elemental de problemas de valor inicial

Las ecuaciones diferenciales se usan para modelar problemas en la ciencia y la ingeniería que implican el cambio de alguna variable respecto a otra. Muchos de estos problemas requieren la solución de un *problema de valor inicial*, es decir, la solución a una ecuación diferencial que satisface una condición inicial determinada.

En situaciones de la vida real común, la ecuación diferencial que modela el problema es demasiado complicada para resolverse de manera exacta y se toma uno de dos enfoques para aproximar la solución. El primer enfoque es modificar el problema al simplificar la ecuación diferencial por una que se pueda resolver de manera exacta y, a continuación, utilizar la solución de la ecuación simplificada para aproximar la solución para el problema original. El otro enfoque, que se examinará en este capítulo, usa métodos para aproximar la solución del problema original. Este es el enfoque que se toma con más frecuencia debido a que los métodos de aproximación dan resultados más precisos e información del error más realista.

Los métodos que consideramos en este capítulo no producen una aproximación continua para la solución del problema de valor inicial. Más bien, las aproximaciones se encuentran en ciertos puntos específicos y a menudo igualmente espaciados. Normalmente se usa algún método de interpolación, por lo general el de Hermite, si se necesitan valores intermedios.

Nosotros necesitamos algunas definiciones y resultados a partir de la teoría de las ecuaciones diferenciales ordinarias, al considerar métodos para aproximar las soluciones a los problemas de valor inicial.

Definición 5.1 Se dice que una función $f(t, y)$ satisface la **condición de Lipschitz** en la variable y en un conjunto $D \subset \mathbb{R}^2$ si existe una constante $L > 0$ con

$$|f(t, y_1) - f(t, y_2,)| \leq L|y_1 - y_2|,$$

siempre que (t, y_1) y (t, y_2) estén en D. La constante L recibe el nombre de **constante de Lipschitz** para f. ■

Ejemplo 1 Muestre que $f(t, y) = t|y|$ satisface la condición de Lipschitz en el intervalo $D = \{(t, y) \mid 1 \leq t \leq 2 \text{ y } -3 \leq y \leq 4\}$.

Solución Para cada par de puntos (t, y_1) y (t, y_2) en D, tenemos

$$|f(t, y_1) - f(t, y_2)| = |t|y_1| - t|y_2|| = |t|\,||y_1| - |y_2|| \leq 2|y_1 - y_2|.$$

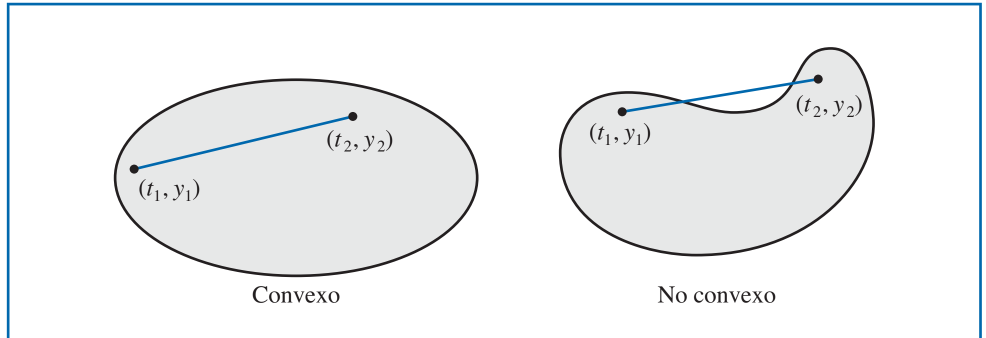
Por lo tanto, f satisface la condición de Lipschitz en D en la variable y con la constante 2 de Lipschitz. El valor más pequeño posible para la constante de Lipschitz para este problema es $L = 2$ porque, por ejemplo,

$$|f(2, 1) - f(2, 0)| = |2 - 0| = 2|1 - 0|.$$ ■

Definición 5.2 Se dice que un conjunto $D \subset \mathbb{R}^2$ es **convexo** siempre que (t_1, y_1) y (t_2, y_2) pertenezcan a D, entonces $((1 - \lambda)t_1 + \lambda t_2, (1 - \lambda)y_1 + \lambda y_2)$ también pertenece a D para cada λ en $[0, 1]$. ■

En términos geométricos, la definición 5.2 establece que un conjunto es convexo a condición de que siempre que dos puntos pertenezcan a un conjunto, todo segmento de línea recta entre los puntos también pertenezca al conjunto (consulte la figura 5.1 y el ejercicio 7.) En general, los conjuntos que consideramos en este capítulo son de la forma $D = \{(t, y) \mid a \leq t \leq b \text{ y } -\infty < y < \infty\}$ para algunas constantes a y b. Es fácil verificar (consulte el ejercicio 9) que estos conjuntos son convexos.

Figura 5.1



Teorema 5.3 Suponga que $f(t, y)$ se define sobre un conjunto convexo $D \subset \mathbb{R}^2$. Si existe una constante $L > 0$ con

$$\left|\frac{\partial f}{\partial y}(t, y)\right| \leq L, \quad \text{para todo } (t, y) \in D, \tag{5.1}$$

entonces f satisface la condición de Lipschitz en D en la variable y con constante L de Lipschitz. ■

Rudolf Lipschitz (1832–1903) trabajó en muchas ramas de las matemáticas, incluyendo la teoría numérica, las series de Fourier, las ecuaciones diferenciales, la mecánica analítica y la teoría del potencial. Es mejor conocido por esta generalización del trabajo de Augustin–Louis Cauchy (1789–1857) y Guiseppe Peano (1856–1932).

La demostración del teorema 5.3 se analiza en el ejercicio 8; es similar a la prueba del resultado correspondiente para funciones de una variable, analizadas en el ejercicio 28 de la sección 1.1.

Como veremos en el siguiente teorema, a menudo es de interés significativo para determinar si la función implicada en el problema de valor inicial satisface la condición de Lipschitz

Definición 5.5 El problema de valor inicial

$$\frac{dy}{dt} = f(t, y), \quad a \leq t \leq b, \quad y(a) = \alpha, \tag{5.2}$$

se dice que es un **problema bien planteado** si:

- Existe una única solución, $y(t)$, y
- Existen constantes $\varepsilon_0 > 0$ y $k > 0$, tales que para cualquier ε en $(0, \varepsilon_0)$, siempre que $\delta(t)$ es continua con $|\delta(t)| < \varepsilon$ para toda t en $[a, b]$, y cuando $|\delta_0| < \varepsilon$, el problema de valor inicial

$$\frac{dz}{dt} = f(t, z) + \delta(t), \quad a \leq t \leq b, \quad z(a) = \alpha + \delta_0, \tag{5.3}$$

tiene una única solución $z(t)$ que satisface

$$|z(t) - y(t)| < k\varepsilon \quad \text{para toda } t \text{ en } [a, b].$$ ■

El problema especificado por la ecuación (5.3) recibe el nombre de **problema perturbado** relacionado con el problema original en la ecuación (5.2). Supone la posibilidad de un error introducido en la declaración de la ecuación diferencial, así como un error ∂_0 presente en la condición inicial.

Los métodos numéricos quizá impliquen resolver un problema perturbado debido a que cualquier error de redondeo introducido en la representación perturba el problema original. A menos que se plantee el problema original, existen pocas razones para esperar que la solución numérica para un problema de este tipo se aproximaría con precisión a la solución del problema original.

El siguiente teorema especifica las condiciones que garantizan un problema de valor inicial bien planteado. La demostración de este teorema se puede encontrar en [BiR], pp. 142–147.

Teorema 5.6 Suponga que $D = \{(t, y) \mid a \leq t \leq b \text{ y } -\infty < y < \infty\}$. Si f es continua y satisface la condición de Lipschitz en la variable y sobre el conjunto D, entonces el problema de valor inicial

$$\frac{dy}{dt} = f(t, y), \quad a \leq t \leq b, \quad y(a) = \alpha$$

está bien planteado. ■

Ejemplo 3 Muestre que el problema de valor inicial

$$\frac{dy}{dt} = y - t^2 + 1, \quad 0 \leq t \leq 2, \quad y(0) = 0.5, \tag{5.4}$$

está bien planteado en $D = \{(t, y) \mid 0 \leq t \leq 2 \text{ y } -\infty < y < \infty\}$.

Solución Puesto que

$$\left|\frac{\partial(y - t^2 + 1)}{\partial y}\right| = |1| = 1,$$

el teorema 5.3 implica que $f(t, y) = y - t^2 + 1$ satisface la condición de Lipschitz en y sobre D con la constante de Lipschitz 1. Puesto que f es continua en D, el teorema 5.6 implica que el problema está bien planteado.

Como ilustración, considere la solución para el problema perturbado

$$\frac{dz}{dt} = z - t^2 + 1 + \delta, \quad 0 \leq t \leq 2, \quad z(0) = 0.5 + \delta_0, \tag{5.5}$$

donde δ y δ_0 son constantes. Las soluciones a las ecuaciones (5.4) y (5.5) son

$$y(t) = (t+1)^2 - 0.5e^t \quad \text{y} \quad z(t) = (t+1)^2 + (\delta + \delta_0 - 0.5)e^t - \delta,$$

respectivamente.

Suponga que ε es un número positivo. Si $|\delta| < \varepsilon$ y $|\delta_0| < \varepsilon$, entonces

$$|y(t) - z(t)| = |(\delta + \delta_0)e^t - \delta| \leq |\delta + \delta_0|e^2 + |\delta| \leq (2e^2 + 1)\varepsilon$$

para todas las t. Esto implica que el problema (5.4) está bien planteado con $k(\varepsilon) = 2e^2 + 1$ para todas las $\varepsilon > 0$. ■

La sección Conjunto de ejercicios 5.1 está disponible en línea. Encuentre la ruta de acceso en las páginas preliminares.

5.2 Método de Euler

El método de Euler es la técnica de aproximación más básica para resolver problemas de valor inicial. A pesar de que rara vez se usa en la práctica, la simplicidad de su derivación se puede utilizar para ilustrar las técnicas relacionadas con la construcción de algunas de las técnicas más avanzadas, sin el álgebra engorrosa que acompaña estas construcciones.

El objetivo del método de Euler es obtener aproximaciones para el problema de valor inicial bien planteado

$$\frac{dy}{dt} = f(t, y), \quad a \leq t \leq b, \quad y(a) = \alpha. \tag{5.6}$$

No se obtendrá una aproximación continua a la solución $y(t)$; en su lugar, las aproximaciones para y se generarán en varios valores, llamados **puntos de malla**, en el intervalo $[a, b]$. Una vez que se obtiene la solución aproximada en los puntos, la solución aproximada en otros puntos en el intervalo se puede encontrar a través de interpolación.

Primero estipulamos que los puntos de malla están igualmente espaciados a lo largo del intervalo $[a, b]$. Esta condición se garantiza al seleccionar un entero positivo N, al establecer $h = (b - a)/N$, y seleccionar los puntos de malla

$$t_i = a + ih, \quad \text{para cada } i = 0, 1, 2, \ldots, N.$$

El uso de métodos de diferencia básica para aproximar la solución a las ecuaciones diferenciales fue uno de los diferentes temas matemáticos que se presentaron primero al público por el más prolífico de los matemáticos, Leonhard Euler (1707–1783).

La distancia común entre los puntos $h = t_{i+1} - t_i$ recibe el nombre de **tamaño de paso**.

Usaremos el teorema de Taylor para deducir el método de Euler. Suponga que $y(t)$, la única solución para (5.6), tiene dos derivadas continuas en $[a, b]$, de tal forma que cada $i = 0, 1, 2, \ldots, N - 1$,

$$y(t_{i+1}) = y(t_i) + (t_{i+1} - t_i)y'(t_i) + \frac{(t_{i+1} - t_i)^2}{2} y''(\xi_i),$$

para algún número ξ_i en (t_i, t_{i+1}). Puesto que $h = t_{i+1} - t_i$, tenemos

$$y(t_{i+1}) = y(t_i) + hy'(t_i) + \frac{h^2}{2} y''(\xi_i),$$

y ya que $y(t)$ satisface la ecuación diferencial (5.6),

$$y(t_{i+1}) = y(t_i) + hf(t_i, y(t_i)) + \frac{h^2}{2} y''(\xi_i). \tag{5.7}$$

El método de Euler construye $w_i \approx y(t_i)$, para cada $i = 1, 2, \ldots, N$, al borrar el término restante. Por lo tanto, el método de Euler es

$$w_0 = \alpha,$$
$$w_{i+1} = w_i + hf(t_i, w_i), \quad \text{para cada } i = 0, 1, \ldots, N-1. \tag{5.8}$$

Ilustración En el ejemplo 1, usaremos un algoritmo para el método de Euler para aproximar la solución de

$$y' = y - t^2 + 1, \quad 0 \le t \le 2, \quad y(0) = 0.5,$$

en $t = 2$. Aquí simplemente ilustraremos los pasos en la técnica cuando tenemos $h = 0.5$. Para este problema $f(t, y) = y - t^2 + 1$; por lo que,

$$w_0 = y(0) = 0.5;$$
$$w_1 = w_0 + 0.5\left(w_0 - (0.0)^2 + 1\right) = 0.5 + 0.5(1.5) = 1.25;$$
$$w_2 = w_1 + 0.5\left(w_1 - (0.5)^2 + 1\right) = 1.25 + 0.5(2.0) = 2.25;$$
$$w_3 = w_2 + 0.5\left(w_2 - (1.0)^2 + 1\right) = 2.25 + 0.5(2.25) = 3.375;$$

y

$$y(2) \approx w_4 = w_3 + 0.5\left(w_3 - (1.5)^2 + 1\right) = 3.375 + 0.5(2.125) = 4.4375.$$ ■

La ecuación (5.8) recibe el nombre de **ecuación de diferencia** relacionada con el método de Euler. Como veremos más adelante en este capítulo, la teoría y la solución de ecuaciones de diferencia son paralelas, en muchas formas, a la teoría y solución de ecuaciones diferenciales. El algoritmo 5.1 implementa el método de Euler.

ALGORITMO 5.1

Método de Euler

Para aproximar la solución del problema de valor inicial

$$y' = f(t, y), \quad a \le t \le b, \quad y(a) = \alpha,$$

en $(N + 1)$ números igualmente espaciados en el intervalo $[a, b]$:

ENTRADA extremos a, b; entero N; condición inicial α.

SALIDA aproximación w para y en $(N + 1)$ valores de t.

Paso 1 Determine $h = (b - a)/N$;
$t = a$;
$w = \alpha$;
SALIDA (t, w).

Paso 2 Para $i = 1, 2, \ldots, N$ haga los pasos 3, 4.

Paso 3 Determine $w = w + hf(t, w)$; (*Calcule* w_i.)
$t = a + ih$. (*Calcule* t_i.)

Paso 4 SALIDA (t, w).

Paso 5 PARE. ■

Para interpretar el método de Euler de manera geométrica, observe que cuando w_i es una aproximación cercana para $y(t_i)$, la suposición de que el problema está bien planteado implica que

$$f(t_i, w_i) \approx y'(t_i) = f(t_i, y(t_i)).$$

La gráfica de la función que resalta $y(t_i)$ se muestra en la figura 5.2. Un paso en el método de Euler aparece en la figura 5.3 y una serie de pasos aparece en la figura 5.4.

Figura 5.2

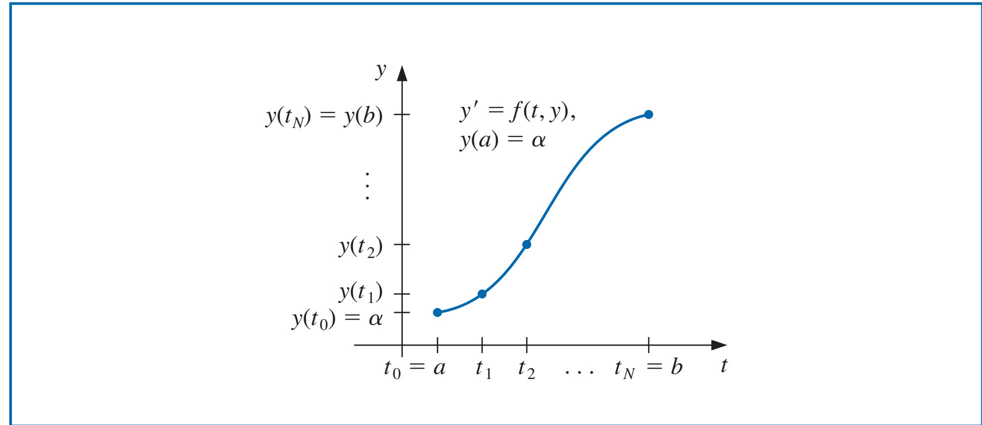


Figura 5.3

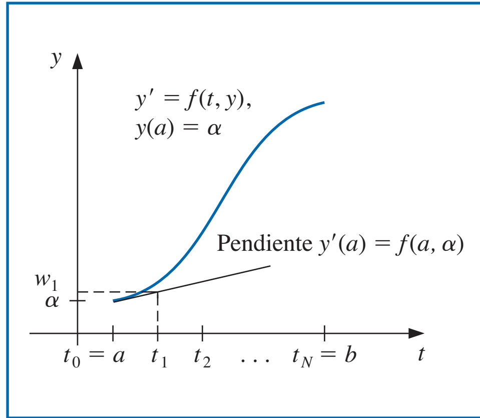
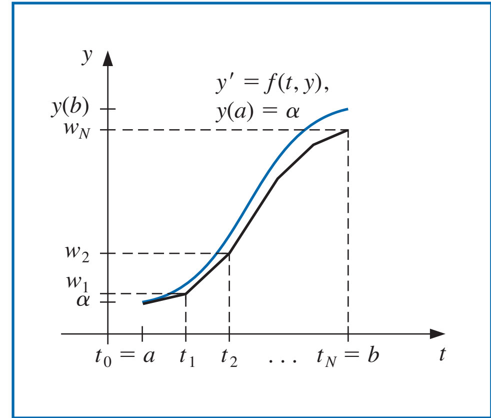


Figura 5.4



Ejemplo 1 El método de Euler se usó en la primera ilustración con $h = 0.5$ para aproximar la solución al problema de valor inicial

$$y' = y - t^2 + 1, \quad 0 \leq t \leq 2, \quad y(0) = 0.5.$$

Utilice el algoritmo 5.1 con $N = 10$ para determinar aproximaciones y compárelas con los valores exactos dados por $y(t) = (t + 1)^2 - 0.5e^t$.

Solución Con $N = 10$, tenemos $h = 0.2$, $t_i = 0.2i$, $w_0 = 0.5$, y

$$w_{i+1} = w_i + h(w_i - t_i^2 + 1) = w_i + 0.2[w_i - 0.04i^2 + 1] = 1.2w_i - 0.008i^2 + 0.2,$$

para $i = 0, 1, \ldots, 9$. Por lo que,

$$w_1 = 1.2(0.5) - 0.008(0)^2 + 0.2 = 0.8, \quad w_2 = 1.2(0.8) - 0.008(1)^2 + 0.2 = 1.152,$$

y así sucesivamente. La tabla 5.1 muestra la comparación entre los valores aproximados en t_i y los valores reales. ■

Tabla 5.1

t_i	w_i	$y_i = y(t_i)$	$\lvert y_i - w_i \rvert$
0.0	0.5000000	0.5000000	0.0000000
0.2	0.8000000	0.8292986	0.0292986
0.4	1.1520000	1.2140877	0.0620877
0.6	1.5504000	1.6489406	0.0985406
0.8	1.9884800	2.1272295	0.1387495
1.0	2.4581760	2.6408591	0.1826831
1.2	2.9498112	3.1799415	0.2301303
1.4	3.4517734	3.7324000	0.2806266
1.6	3.9501281	4.2834838	0.3333557
1.8	4.4281538	4.8151763	0.3870225
2.0	4.8657845	5.3054720	0.4396874

Observe que el error crece ligeramente conforme el valor de t aumenta. Este crecimiento de error controlado es una consecuencia de la estabilidad del método de Euler, lo cual implica que se espera que el error no crezca de una manera mejor a la forma lineal.

Cotas del error para el método de Euler

A pesar de que el método de Euler no es por completo apropiado para garantizar su uso en la práctica, es suficientemente básico para analizar el error producido a partir de esta aplicación. El análisis de error para los métodos más precisos que consideramos en las secciones subsiguientes sigue el mismo patrón, pero es más complicado.

Para derivar una cota del error para el método de Euler, necesitamos dos lemas de cálculo.

Lema 5.7 Para toda $x \geq -1$ y cualquier m positiva, tenemos $0 \leq (1+x)^m \leq e^{mx}$.

Demostración Al aplicar el teorema de Taylor con $f(x) = e^x$, $x_0 = 0$, y $n = 1$ obtenemos

$$e^x = 1 + x + \frac{1}{2}x^2 e^{\xi},$$

donde ξ está entre x y cero. Por lo tanto,

$$0 \leq 1 + x \leq 1 + x + \frac{1}{2}x^2 e^{\xi} = e^x,$$

y, puesto que $1 + x \geq 0$, tenemos

$$0 \leq (1+x)^m \leq (e^x)^m = e^{mx}.$$

■

Lema 5.8 Si s y t son números reales positivos, $\{a_i\}_{i=0}^{k}$ es una sucesión que satisface $a_0 \geq -t/s$, y

$$a_{i+1} \leq (1+s)a_i + t, \quad \text{para cada } i = 0, 1, 2, \ldots, k-1, \tag{5.9}$$

entonces

$$a_{i+1} \leq e^{(i+1)s}\left(a_0 + \frac{t}{s}\right) - \frac{t}{s}.$$

Demostración Para un entero fijo i, la desigualdad (5.9) implica que

$$\begin{aligned}
a_{i+1} &\leq (1+s)a_i + t \\
&\leq (1+s)[(1+s)a_{i-1} + t] + t = (1+s)^2 a_{i-1} + [1 + (1+s)]t \\
&\leq (1+s)^3 a_{i-2} + \left[1 + (1+s) + (1+s)^2\right] t \\
&\ \ \vdots \\
&\leq (1+s)^{i+1} a_0 + \left[1 + (1+s) + (1+s)^2 + \cdots + (1+s)^i\right] t.
\end{aligned}$$

Pero

$$1 + (1+s) + (1+s)^2 + \cdots + (1+s)^i = \sum_{j=0}^{i} (1+s)^j$$

es una serie geométrica con radio $(1 + s)$ que se suma a

$$\frac{1-(1+s)^{i+1}}{1-(1+s)} = \frac{1}{s}[(1+s)^{i+1} - 1].$$

Por lo tanto,

$$a_{i+1} \le (1+s)^{i+1}a_0 + \frac{(1+s)^{i+1}-1}{s}t = (1+s)^{i+1}\left(a_0 + \frac{t}{s}\right) - \frac{t}{s},$$

y por el lema 5.7 con $x = 1 + s$ obtenemos

$$a_{i+1} \le e^{(i+1)s}\left(a_0 + \frac{t}{s}\right) - \frac{t}{s}.$$

■

Teorema 5.9 Suponga que f es continua y satisface la condición de Lipschitz con constante L en

$$D = \{\, (t, y) \mid a \le t \le b \text{ y } -\infty < y < \infty \,\}$$

y que existe una constante M con

$$|y''(t)| \le M, \quad \text{para todas las } t \in [a, b],$$

donde $y(t)$ denota la única solución para el problema de valor inicial

$$y' = f(t, y), \quad a \le t \le b, \quad y(a) = \alpha.$$

Sean $w_0, w_1, \ldots, w_N$ las aproximaciones generadas por el método de Euler para un entero positivo N. Entonces, para cada $i = 0, 1, 2, \ldots, N$,

$$|y(t_i) - w_i| \le \frac{hM}{2L}\left[e^{L(t_i - a)} - 1\right]. \tag{5.10}$$

Demostración Cuando $i = 0$, el resultado es claramente verdadero ya que $y(t_0) = w_0 = \alpha$.
A partir de la ecuación (5.7), tenemos

$$y(t_{i+1}) = y(t_i) + hf(t_i, y(t_i)) + \frac{h^2}{2}y''(\xi_i),$$

para $i = 0, 1, \ldots, N-1$, y a partir de las ecuaciones en (5.8)

$$w_{i+1} = w_i + hf(t_i, w_i).$$

Usando la notación $y_i = y(t_i)$ y $y_{i+1} = y(t_{i+1})$, restamos estas dos ecuaciones para obtener

$$y_{i+1} - w_{i+1} = y_i - w_i + h[f(t_i, y_i) - f(t_i, w_i)] + \frac{h^2}{2}y''(\xi_i)$$

Por lo tanto,

$$|y_{i+1} - w_{i+1}| \le |y_i - w_i| + h|f(t_i, y_i) - f(t_i, w_i)| + \frac{h^2}{2}|y''(\xi_i)|.$$

Ahora f satisface la condición de Lipschitz en la segunda variable con constante L y $|y''(t)| \leq M$, por lo que

$$|y_{i+1} - w_{i+1}| \leq (1 + hL)|y_i - w_i| + \frac{h^2 M}{2}.$$

De acuerdo con el lema 5.8 y haciendo $s = hL, t = h^2 M/2$, y $a_j = |y_j - w_j|$, para cada $j = 0, 1, \ldots, N$, observaremos que

$$|y_{i+1} - w_{i+1}| \leq e^{(i+1)hL}\left(|y_0 - w_0| + \frac{h^2 M}{2hL}\right) - \frac{h^2 M}{2hL}.$$

Puesto que $|y_0 - w_0| = 0$ y $(i + 1)h = t_{i+1} - t_0 = t_{i+1} - a$, esto implica que

$$|y_{i+1} - w_{i+1}| \leq \frac{hM}{2L}(e^{(t_{i+1}-a)L} - 1),$$

para cada $i = 0, 1, \ldots, N - 1$. ■

La debilidad del teorema 5.9 depende del requisito de conocer una cota para la segunda derivada de la solución. A pesar de que, a menudo, esta condición nos prohíbe obtener una cota de error realista, se debería observar que si existe $\partial f/\partial t$ y $\partial f/\partial y$, la regla de la cadena para la diferenciación parcial implica que

$$y''(t) = \frac{dy'}{dt}(t) = \frac{df}{dt}(t, y(t)) = \frac{\partial f}{\partial t}(t, y(t)) + \frac{\partial f}{\partial y}(t, y(t)) \cdot f(t, y(t)).$$

Por lo tanto, algunas veces es posible obtener una cota de error para $y''(t)$ sin conocer explícitamente $y(t)$.

Ejemplo 2 La solución para el problema de valor inicial

$$y' = y - t^2 + 1, \quad 0 \leq t \leq 2, \quad y(0) = 0.5,$$

se aproximó en el ejemplo 1 con el método de Euler con $h = 0.2$. Utilice la desigualdad en el teorema 5.9 para encontrar una cota para los errores de aproximación y compárelos con los errores reales.

Solución Puesto que $f(t, y) = y - t^2 + 1$, tenemos $\partial f(t, y)/\partial y = 1$ para todas las y, por lo que $L = 1$. Para este problema, la solución exacta es $y(t) = (t + 1)^2 - 0.5e^t$, por lo que $y''(t) = 2 - 0.5e^t$ y

$$|y''(t)| \leq 0.5e^2 - 2, \quad \text{para todas las } t \in [0, 2].$$

Por medio de desigualdad en la cota de error para el método de Euler con $h = 0.2, L = 1$, y $M = 0.5e^2 - 2$ da

$$|y_i - w_i| \leq 0.1(0.5e^2 - 2)(e^{t_i} - 1).$$

Por lo tanto

$$|y(0.2) - w_1| \leq 0.1(0.5e^2 - 2)(e^{0.2} - 1) = 0.03752,$$

$$|y(0.4) - w_2| \leq 0.1(0.5e^2 - 2)(e^{0.4} - 1) = 0.08334,$$

y así sucesivamente. La tabla 5.2 enumera el error real encontrado en el ejemplo 1, junto con la cota de error. Observe que aunque se usó la cota verdadera para la segunda derivada de la solución, la cota de error es considerablemente superior que el error real, en especial para los valores mayores de t. ■

Tabla 5.2

t_i	0.2	0.4	0.6	0.8	1.0	1.2	1.4	1.6	1.8	2.0
Error real	0.02930	0.06209	0.09854	0.13875	0.18268	0.23013	0.28063	0.33336	0.38702	0.43969
Cota de error	0.03752	0.08334	0.13931	0.20767	0.29117	0.39315	0.51771	0.66985	0.85568	1.08264

La principal importancia de la fórmula de la cota de error determinada en el teorema 5.9 es que la cota depende linealmente del tamaño de paso h. Por consiguiente, disminuir el tamaño de paso debería proporcionar mayor precisión para las aproximaciones en la misma medida.

Olvidado en el resultado del teorema 5.9 está el efecto que el error de redondeo representa en la selección del tamaño de paso. Conforme h se vuelve más pequeño, se necesitan más cálculos y se espera más error de redondeo. Entonces, en la actualidad, la forma de ecuación de diferencia

$$w_0 = \alpha,$$
$$w_{i+1} = w_i + hf(t_i, w_i), \quad \text{para cada } i = 0, 1, \ldots, N-1,$$

no se utiliza para calcular la aproximación a la solución y_i en un punto de malla t_i. En su lugar, usamos una ecuación de la forma

$$u_0 = \alpha + \delta_0,$$
$$u_{i+1} = u_i + hf(t_i, u_i) + \delta_{i+1}, \quad \text{para cada } i = 0, 1, \ldots, N-1, \tag{5.11}$$

donde δ_i denota el error de redondeo asociado con u_i. Al utilizar métodos similares a aquellos en la prueba del teorema 5.9 podemos producir una cota de error para las aproximaciones de dígitos finitos para y_i provistos por el método de Euler.

Teorema 5.10 Si $y(t)$ denota la única solución para el problema de valor inicial

$$y' = f(t, y), \quad a \le t \le b, \quad y(a) = \alpha, \tag{5.12}$$

y $u_0, u_1, \ldots, u_N$ son las aproximaciones obtenidas de la ecuación (5.11). Si $|\delta_i| < \delta$ para cada $i = 0, 1, \ldots, N$ y la hipótesis del teorema 5.9 son aplicables a la ecuación (5.12), entonces

$$|y(t_i) - u_i| \le \frac{1}{L}\left(\frac{hM}{2} + \frac{\delta}{h}\right)[e^{L(t_i-a)} - 1] + |\delta_0|e^{L(t_i-a)}, \tag{5.13}$$

para cada $i = 0, 1, \ldots, N$. ■

La cota de error (5.13) ya no es lineal en h. De hecho, puesto que

$$\lim_{h\to 0}\left(\frac{hM}{2} + \frac{\delta}{h}\right) = \infty,$$

se esperaría que el error se vuelva más grande para los valores suficientemente pequeños de h. El cálculo se puede usar para determinar una cota inferior para el tamaño de paso h. Si $E(h) = (hM/2) + (\delta/h)$ implica que $E'(h) = (M/2) - (\delta/h^2)$:

Si $h < \sqrt{2\delta/M}$, entonces $E'(h) < 0$ y $E(h)$ disminuye.

Si $h > \sqrt{2\delta/M}$, entonces $E'(h) > 0$ y $E(h)$ aumenta.

El valor mínimo de $E(h)$ se presenta cuando

$$h = \sqrt{\frac{2\delta}{M}}. \tag{5.14}$$

La disminución de h más allá de este valor tiende a incrementar el error total en la aproximación. Por lo general, sin embargo, el valor de δ es suficientemente pequeño para que esta cota inferior para h no afecte la operación del método de Euler.

La sección Conjunto de ejercicios 5.2 está disponible en línea. Encuentre la ruta de acceso en las páginas preliminares.

5.3 Métodos de Taylor de orden superior

Puesto que el objetivo de una técnica numérica es determinar aproximaciones precisas con mínimo esfuerzo, necesitamos medios para comparar la eficiencia de los diferentes métodos de aproximación. El primer dispositivo que consideramos recibe el nombre de *error de truncamiento local* del método.

El error de truncamiento local en un paso específico mide la cantidad por la que la solución exacta para la ecuación diferencial falla en cuanto a satisfacer la ecuación de diferencia que se usa para la aproximación en ese paso. Esto podría parecer una forma poco probable de comparar el error de varios métodos. En realidad queremos saber qué tan bien satisfacen las aproximaciones generadas con los métodos la ecuación diferencial, no al revés. Sin embargo, no conocemos la solución exacta por lo que, en general, no podemos determinarlo y el error de truncamiento local servirá de manera adecuada para determinar no sólo el error local de un método, sino también el error de aproximación real.

Considere el problema de valor inicial

$$y' = f(t, y), \quad a \leq t \leq b, \quad y(a) = \alpha.$$

Definición 5.11 El método de diferencia

$$w_0 = \alpha$$

$$w_{i+1} = w_i + h\phi(t_i, w_i), \quad \text{para cada } i = 0, 1, \ldots, N-1,$$

tiene **error de truncamiento local**

$$\tau_{i+1}(h) = \frac{y_{i+1} - (y_i + h\phi(t_i, y_i))}{h} = \frac{y_{i+1} - y_i}{h} - \phi(t_i, y_i),$$

para cada $i = 0, 1, \ldots, N-1$, donde y_i y y_{i+1} denotan la solución de la ecuación diferencial en t_i y t_{i+1}, respectivamente. ■

Por ejemplo, el método de Euler tiene error de truncamiento en el i-ésimo paso

$$\tau_{i+1}(h) = \frac{y_{i+1} - y_i}{h} - f(t_i, y_i), \quad \text{para cada } i = 0, 1, \ldots, N-1.$$

Este error es un *error local* porque mide la precisión del método en un paso específico, al suponer que el método era exacto en el paso anterior. Como tal, depende de la ecuación diferencial, del tamaño de paso y del paso particular en la aproximación.

Al considerar la ecuación (5.7) en la sección previa, observamos que el método de Euler tiene

$$\tau_{i+1}(h) = \frac{h}{2}y''(\xi_i), \quad \text{para algunas } \xi_i \text{ en } (t_i, t_{i+1}).$$

Cuando se sabe que $y''(t)$ está acotada por una constante M en $[a, b]$, esto implica

$$|\tau_{i+1}(h)| \leq \frac{h}{2}M,$$

por lo que el error de truncamiento local es $O(h)$.

Una forma de seleccionar métodos de ecuación de diferencia para resolver ecuaciones diferenciales ordinarias de tal forma que sus errores de truncamiento local son $O(h^p)$ para un valor de p tan grande como sea posible, mientras se mantienen el número y la complejidad de los cálculos de los métodos dentro de un límite razonable.

Puesto que el método de Euler se derivó del teorema de Taylor con $n = 1$ para aproximar la solución de la ecuación diferencial, nuestro primer intento para encontrar métodos para mejorar las propiedades de convergencia de los métodos de diferencia es ampliar esta técnica de derivación a valores más grandes de n.

Los métodos en esta sección utilizan polinomios de Taylor y el conocimiento de la derivada en un nodo para aproximar el valor de la función en un nodo nuevo.

Suponga que la solución $y(t)$ para el problema de valor inicial

$$y' = f(t, y), \quad a \le t \le b, \quad y(a) = \alpha,$$

tiene $(n + 1)$ derivadas continuas. Si ampliamos la solución, $y(t)$, en términos de su enésimo polinomio de Taylor alrededor de t_i y se evalúan en t_{i+1}, obtenemos

$$y(t_{i+1}) = y(t_i) + hy'(t_i) + \frac{h^2}{2}y''(t_i) + \cdots + \frac{h^n}{n!}y^{(n)}(t_i) + \frac{h^{n+1}}{(n+1)!}y^{(n+1)}(\xi_i), \quad (5.15)$$

para algunas ξ_i en (t_i, t_{i+1}).

La diferenciación sucesiva de la solución, $y(t)$, da

$$y'(t) = f(t, y(t)), \quad y''(t) = f'(t, y(t)), \quad \text{y, en general,} \quad y^{(k)}(t) = f^{(k-1)}(t, y(t)).$$

Al sustituir estos resultados en la ecuación (5.15) obtenemos

$$\begin{aligned} y(t_{i+1}) = y(t_i) + hf(t_i, y(t_i)) + \frac{h^2}{2}f'(t_i, y(t_i)) + \cdots \\ + \frac{h^n}{n!}f^{(n-1)}(t_i, y(t_i)) + \frac{h^{n+1}}{(n+1)!}f^{(n)}(\xi_i, y(\xi_i)). \end{aligned} \quad (5.16)$$

El método de ecuación de diferencia correspondiente a la ecuación (5.16) se obtiene al borrar el término restante relacionado con ξ_i.

Método de Taylor de orden *n*

$$\begin{aligned} w_0 &= \alpha, \\ w_{i+1} &= w_i + hT^{(n)}(t_i, w_i), \quad \text{para cada } i = 0, 1, \ldots, N-1, \end{aligned} \quad (5.17)$$

donde

$$T^{(n)}(t_i, w_i) = f(t_i, w_i) + \frac{h}{2}f'(t_i, w_i) + \cdots + \frac{h^{n-1}}{n!}f^{(n-1)}(t_i, w_i).$$

El método de Euler es un método de Taylor de orden uno.

Ejemplo 1 Aplique el método de Taylor de órdenes **a**) 2 y **b**) 4 con $N = 10$ al problema de valor inicial

$$y' = y - t^2 + 1, \quad 0 \le t \le 2, \quad y(0) = 0.5.$$

Solución **a**) Para el método de orden 2, necesitamos la primera derivada de $f(t, y(t)) = y(t) - t^2 + 1$ respecto a la variable t. Puesto que $y' = y - t^2 + 1$, tenemos

$$f'(t, y(t)) = \frac{d}{dt}(y - t^2 + 1) = y' - 2t = y - t^2 + 1 - 2t,$$

por lo que

$$T^{(2)}(t_i, w_i) = f(t_i, w_i) + \frac{h}{2} f'(t_i, w_i) = w_i - t_i^2 + 1 + \frac{h}{2}(w_i - t_i^2 + 1 - 2t_i)$$
$$= \left(1 + \frac{h}{2}\right)(w_i - t_i^2 + 1) - ht_i.$$

Puesto que $N = 10$, tenemos $h = 0.2$, y $t_i = 0.2i$ para cada $i = 1, 2, \ldots, 10$. Por lo tanto, el método de segundo orden se vuelve

$$w_0 = 0.5,$$
$$w_{i+1} = w_i + h\left[\left(1 + \frac{h}{2}\right)(w_i - t_i^2 + 1) - ht_i\right]$$
$$= w_i + 0.2\left[\left(1 + \frac{0.2}{2}\right)(w_i - 0.04i^2 + 1) - 0.04i\right]$$
$$= 1.22w_i - 0.0088i^2 - 0.008i + 0.22.$$

Los primeros dos pasos dan las aproximaciones

$$y(0.2) \approx w_1 = 1.22(0.5) - 0.0088(0)^2 - 0.008(0) + 0.22 = 0.83$$

y

$$y(0.4) \approx w_2 = 1.22(0.83) - 0.0088(0.2)^2 - 0.008(0.2) + 0.22 = 1.2158.$$

Todas las aproximaciones y sus errores se muestran en la tabla 5.3.

Tabla 5.3

t_i	Orden 2 de Taylor w_i	Error $\lvert y(t_i) - w_i \rvert$
0.0	0.500000	0
0.2	0.830000	0.000701
0.4	1.215800	0.001712
0.6	1.652076	0.003135
0.8	2.132333	0.005103
1.0	2.648646	0.007787
1.2	3.191348	0.011407
1.4	3.748645	0.016245
1.6	4.306146	0.022663
1.8	4.846299	0.031122
2.0	5.347684	0.042212

b) Para el método de Taylor de orden 4 necesitamos las primeras tres derivadas de $f(t, y(t))$ respecto a t. De nuevo, por medio de $y' = y - t^2 + 1$, tenemos

$$f'(t, y(t)) = y - t^2 + 1 - 2t,$$
$$f''(t, y(t)) = \frac{d}{dt}(y - t^2 + 1 - 2t) = y' - 2t - 2$$
$$= y - t^2 + 1 - 2t - 2 = y - t^2 - 2t - 1,$$

y

$$f'''(t, y(t)) = \frac{d}{dt}(y - t^2 - 2t - 1) = y' - 2t - 2 = y - t^2 - 2t - 1,$$

por lo que

$$T^{(4)}(t_i, w_i) = f(t_i, w_i) + \frac{h}{2} f'(t_i, w_i) + \frac{h^2}{6} f''(t_i, w_i) + \frac{h^3}{24} f'''(t_i, w_i)$$
$$= w_i - t_i^2 + 1 + \frac{h}{2}(w_i - t_i^2 + 1 - 2t_i) + \frac{h^2}{6}(w_i - t_i^2 - 2t_i - 1)$$
$$+ \frac{h^3}{24}(w_i - t_i^2 - 2t_i - 1)$$
$$= \left(1 + \frac{h}{2} + \frac{h^2}{6} + \frac{h^3}{24}\right)(w_i - t_i^2) - \left(1 + \frac{h}{3} + \frac{h^2}{12}\right)(ht_i)$$
$$+ 1 + \frac{h}{2} - \frac{h^2}{6} - \frac{h^3}{24}.$$

Por lo tanto, el método de Taylor de orden 4 es

$$w_0 = 0.5,$$
$$w_{i+1} = w_i + h\left[\left(1 + \frac{h}{2} + \frac{h^2}{6} + \frac{h^3}{24}\right)(w_i - t_i^2) - \left(1 + \frac{h}{3} + \frac{h^2}{12}\right)ht_i + 1 + \frac{h}{2} - \frac{h^2}{6} - \frac{h^3}{24}\right],$$

para $i = 0, 1, \ldots, N-1$.

Puesto que $N = 10$ y $h = 0.2$, el método se vuelve

$$w_{i+1} = w_i + 0.2\left[\left(1 + \frac{0.2}{2} + \frac{0.04}{6} + \frac{0.008}{24}\right)(w_i - 0.04i^2) - \left(1 + \frac{0.2}{3} + \frac{0.04}{12}\right)(0.04i) + 1 + \frac{0.2}{2} - \frac{0.04}{6} - \frac{0.008}{24}\right]$$
$$= 1.2214w_i - 0.008856i^2 - 0.00856i + 0.2186,$$

para cada $i = 0, 1, \ldots, 9$. Los primeros dos pasos proporcionan las aproximaciones

$$y(0.2) \approx w_1 = 1.2214(0.5) - 0.008856(0)^2 - 0.00856(0) + 0.2186 = 0.8293$$

y

$$y(0.4) \approx w_2 = 1.2214(0.8293) - 0.008856(0.2)^2 - 0.00856(0.2) + 0.2186 = 1.214091.$$

Todas las aproximaciones y sus errores se muestran en la tabla 5.4.

Compare estos resultados con los del método de Taylor de orden 2 en la tabla 5.3 y observará que los resultados de cuarto orden son inmensamente superiores. ■

Tabla 5.4

t_i	Orden 4 de Taylor w_i	Error $\lvert y(t_i) - w_i \rvert$
0.0	0.500000	0
0.2	0.829300	0.000001
0.4	1.214091	0.000003
0.6	1.648947	0.000006
0.8	2.127240	0.000010
1.0	2.640874	0.000015
1.2	3.179964	0.000023
1.4	3.732432	0.000032
1.6	4.283529	0.000045
1.8	4.815238	0.000062
2.0	5.305555	0.000083

Los datos de la tabla 5.4 indican que los resultados del método de Taylor de orden 4 son bastante precisos en los nodos 0.2, 0.4 y así sucesivamente. Pero suponga que necesitamos determinar una aproximación para un punto intermedio en la tabla, por ejemplo, en $t = 1.25$. Si usamos interpolación lineal sobre el método de Taylor de orden 4 para aproximaciones en $t = 1.2$ y $t = 1.4$, tenemos

$$y(1.25) \approx \left(\frac{1.25 - 1.4}{1.2 - 1.4}\right)3.1799640 + \left(\frac{1.25 - 1.2}{1.4 - 1.2}\right)3.7324321 = 3.3180810.$$

El verdadero valor es $y(1.25) = 3.3173285$, por lo que esta aproximación tiene un error de 0.0007525, lo cual es alrededor de 30 veces el promedio de los errores de aproximación en 1.2 y 1.4.

Podemos mejorar en forma significativa la aproximación por medio de la interpolación cúbica de Hermite. La determinación de esta aproximación para $y(1.25)$ requiere aproximaciones para $y'(1.2)$ y $y'(1.4)$, así como para $y(1.2)$ y $y(1.4)$. Sin embargo, las aproximaciones para $y(1.2)$ y $y(1.4)$ están en la tabla y las aproximaciones de la derivada están disponibles a partir de la ecuación diferencial ya que $y'(t) = f(t, y(t))$. En nuestro ejemplo, $y'(t) = y(t) - t^2 + 1$, por lo que

La interpolación de Hermite requiere tanto el valor de la función como su derivada en cada nodo. Esto crea un método de interpolación natural para aproximar ecuaciones diferenciales ya que estos datos están disponibles.

$$y'(1.2) = y(1.2) - (1.2)^2 + 1 \approx 3.1799640 - 1.44 + 1 = 2.7399640$$

y

$$y'(1.4) = y(1.4) - (1.4)^2 + 1 \approx 3.7324327 - 1.96 + 1 = 2.7724321.$$

El procedimiento de diferencia dividida en la sección 3.4 provee la información en la tabla 5.5. Las entradas subrayadas provienen de los datos y las otras entradas utilizan fórmulas de diferencia dividida.

Tabla 5.5

1.2	3.1799640			
		2.7399640		
1.2	3.1799640		0.1118825	
		2.7623405		−0.3071225
1.4	3.7324321		0.0504580	
		2.7724321		
1.4	3.7324321			

El polinomio cúbico de Hermite es

$$y(t) \approx 3.1799640 + (t - 1.2)2.7399640 + (t - 1.2)^2 0.1118825 + (t - 1.2)^2(t - 1.4)(-0.3071225),$$

por lo que

$$y(1.25) \approx 3.1799640 + 0.1369982 + 0.0002797 + 0.0001152 = 3.3173571,$$

un resultado preciso dentro de 0.0000286. Esto es aproximadamente el promedio de los errores en 1.2 y 1.4 y sólo 4% del error obtenido mediante interpolación lineal. Esta mejora de la precisión ciertamente justifica los cálculos adicionales requeridos para el método de Hermite.

Teorema 5.12 Si se usa el método de Taylor de orden n para aproximar las solución de

$$y'(t) = f(t, y(t)), \quad a \le t \le b, \quad y(a) = \alpha,$$

con tamaño de paso h y si $y \in C^{n+1}[a, b]$, entonces el error de truncamiento local es $O(h^n)$.

Demostración Observe que la ecuación (5.16) en la página 206 se puede reescribir como

$$y_{i+1} - y_i - hf(t_i, y_i) - \frac{h^2}{2}f'(t_i, y_i) - \cdots - \frac{h^n}{n!}f^{(n-1)}(t_i, y_i) = \frac{h^{n+1}}{(n+1)!}f^{(n)}(\xi_i, y(\xi_i)),$$

para algunas ξ_i en (t_i, t_{i+1}). Por lo que el error de truncamiento es

$$\tau_{i+1}(h) = \frac{y_{i+1} - y_i}{h} - T^{(n)}(t_i, y_i) = \frac{h^n}{(n+1)!}f^{(n)}(\xi_i, y(\xi_i)),$$

para cada $i = 0, 1, \ldots, N - 1$. Puesto que $y \in C^{n+1}[a, b]$, tenemos $y^{(n+1)}(t) = f^{(n)}(t, y(t))$, está acotada en $[a, b]$ y $\tau_i(h) = O(h^n)$, para cada $i = 1, 2, \ldots, N$. ■

La sección Conjunto de ejercicios 5.3 está disponible en línea. Encuentre la ruta de acceso en las paginas preliminares.

5.4 Método Runge-Kutta

Los métodos de Taylor descritos en la sección anterior tienen la propiedad deseable de error de truncamiento local de orden superior, pero la desventaja de requerir el cálculo y la evaluación de las derivadas de $f(t, y)$. Este es un procedimiento complicado y que toma mucho tiempo para la mayoría de los problemas, por lo que los métodos de Taylor rara vez se usan en la práctica.

A finales de la década de 1800, Carl Runge (1856–1927) utilizó métodos similares a los que se han usado en esta sección para derivar varias fórmulas para aproximar la solución de problemas de valor inicial.

Los **métodos Runge-Kutta** tienen el error de truncamiento local de orden superior a los métodos de Taylor, pero eliminan la necesidad de calcular y evaluar las derivadas de $f(t, y)$. Antes de presentar las ideas detrás de su derivación, necesitamos considerar el teorema de Taylor en dos variables. La prueba de este resultado se puede encontrar en cualquier libro estándar sobre cálculo avanzado (consulte, por ejemplo, [Fu], p. 331).

Teorema 5.13 Suponga que $f(t, y)$ y todas sus derivadas parciales de orden menor o igual a $n + 1$ son continuas en $D = \{ (t, y) \mid a \leq t \leq b, c \leq y \leq d \}$ y si $(t_0, y_0) \in D$. Para cada $(t, y) \in D$, existe ξ entre t y t_0 y μ entre y y y_0 con

En 1901, Martin Wilhem Kutta (1867–1944) generalizó los métodos que Runge desarrolló en 1895 para incorporar sistemas de ecuaciones diferenciales de primer orden. Estas técnicas difieren ligeramente de lo que en la actualidad llamamos métodos de Runge-Kutta.

$$f(t, y) = P_n(t, y) + R_n(t, y),$$

donde

$$\begin{aligned} P_n(t, y) = {} & f(t_0, y_0) + \left[(t - t_0)\frac{\partial f}{\partial t}(t_0, y_0) + (y - y_0)\frac{\partial f}{\partial y}(t_0, y_0) \right] \\ & + \left[\frac{(t - t_0)^2}{2}\frac{\partial^2 f}{\partial t^2}(t_0, y_0) + (t - t_0)(y - y_0)\frac{\partial^2 f}{\partial t \partial y}(t_0, y_0) \right. \\ & \left. + \frac{(y - y_0)^2}{2}\frac{\partial^2 f}{\partial y^2}(t_0, y_0) \right] + \cdots \\ & + \left[\frac{1}{n!}\sum_{j=0}^{n} \binom{n}{j} (t - t_0)^{n-j}(y - y_0)^j \frac{\partial^n f}{\partial t^{n-j}\partial y^j}(t_0, y_0) \right] \end{aligned}$$

y

$$R_n(t, y) = \frac{1}{(n + 1)!}\sum_{j=0}^{n+1} \binom{n + 1}{j} (t - t_0)^{n+1-j}(y - y_0)^j \frac{\partial^{n+1} f}{\partial t^{n+1-j}\partial y^j}(\xi, \mu).$$ ■

La función $P_n(t, y)$ recibe el nombre del **enésimo polinomio de Taylor en dos variables** para la función f cerca de (t_0, y_0), y $R_n(t, y)$ es el término restante asociado con $P_n(t, y)$.

Ejemplo 1 Determine $P_2(t, y)$, el segundo polinomio de Taylor cerca de $(2, 3)$ para la función

$$f(t, y) = \exp\left[-\frac{(t - 2)^2}{4} - \frac{(y - 3)^2}{4} \right] \cos(2t + y - 7).$$

Solución Para determinar $P_2(t, y)$, necesitamos los valores de f y su primera y segunda derivadas parciales en $(2, 3)$. Tenemos lo siguiente

$$f(t, y) = \exp\left[-\frac{(t - 2)^2}{4} \right] \exp\left[-\frac{(y - 3)^2}{4} \right] \cos(2(t - 2) + (y - 3))$$

$$f(2, 3) = e^{(-0^2/4 - 0^2/4)} \cos(4 + 3 - 7) = 1,$$

$$\begin{aligned} \frac{\partial f}{\partial t}(t, y) = {} & \exp\left[-\frac{(t - 2)^2}{4} \right] \exp\left[-\frac{(y - 3)^2}{4} \right] \left[\frac{1}{2}(t - 2)\cos(2(t - 2) + (y - 3)) \right. \\ & \left. + \frac{1}{2}(\text{sen}(2(t - 2) + (y - 3)) \right] \end{aligned}$$

$$\frac{\partial f}{\partial t}(2, 3) = 0,$$

$$\begin{aligned} \frac{\partial f}{\partial y}(t, y) = {} & \exp\left[-\frac{(t - 2)^2}{4} \right] \exp\left[-\frac{(y - 3)^2}{4} \right] \left[\frac{1}{2}(y - 3)\cos(2(t - 2) \right. \\ & \left. + (y - 3)) + \text{sen}(2(t - 2) + (y - 3)) \right] \end{aligned}$$

$$\frac{\partial f}{\partial y}(2, 3) = 0,$$

$$\frac{\partial^2 f}{\partial t^2}(t, y) = \exp\left[-\frac{(t-2)^2}{4}\right]\exp\left[-\frac{(y-3)^2}{4}\right]\left[\left(-\frac{9}{2}+\frac{(t-2)^2}{4}\right)\right.$$
$$\left.\times\cos(2(t-2)+(y-3)) + 2(t-2)\,\text{sen}(2(t-2)+(y-3))\right]$$

$$\frac{\partial^2 f}{\partial t^2}(2, 3) = -\frac{9}{2},$$

$$\frac{\partial^2 f}{\partial y^2}(t, y) = \exp\left[-\frac{(t-2)^2}{4}\right]\exp\left[-\frac{(y-3)^2}{4}\right]\left[\left(-\frac{3}{2}+\frac{(y-3)^2}{4}\right)\right.$$
$$\left.\times\cos(2(t-2)+(y-3)) + (y-3)\,\text{sen}(2(t-2)+(y-3))\right]$$

$$\frac{\partial^2 f}{\partial y^2}(2, 3) = -\frac{3}{2},$$

y

$$\frac{\partial^2 f}{\partial t \partial y}(t, y) = \exp\left[-\frac{(t-2)^2}{4}\right]\exp\left[-\frac{(y-3)^2}{4}\right]\left[\left(-2+\frac{(t-2)(y-3)}{4}\right.\right.$$
$$\left.\left.\times\cos(2(t-2)+(y-3)) + \left(\frac{(t-2)}{2}+(y-3)\right)\text{sen}(2(t-2)+(y-3))\right)\right]$$

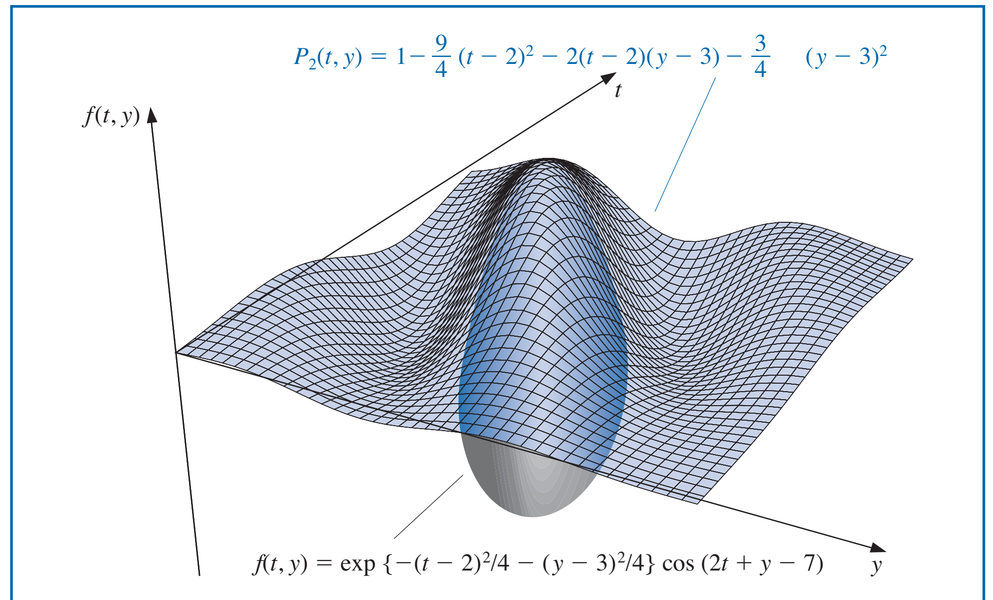
$$\frac{\partial^2 f}{\partial t \partial y}(2, 3) = -2.$$

Por lo que,

$$P_2(t, y) = f(2,3) + \left[(t-2)\frac{\partial f}{\partial t}(2,3) + (y-3)\frac{\partial f}{\partial y}(2,3)\right] + \left[\frac{(t-2)^2}{2}\frac{\partial^2 f}{\partial t^2}(2,3)\right.$$
$$\left. + (t-2)(y-3)\frac{\partial^2 f}{\partial t \partial y}(2,3) + \frac{(y-3)}{2}\frac{\partial^2 f}{\partial y^2}(2,3)\right]$$
$$= 1 - \frac{9}{4}(t-2)^2 - 2(t-2)(y-3) - \frac{3}{4}(y-3)^2.$$
■

Una ilustración de la precisión de $P_2(t, y)$ cerca de (2,3) se observa en la figura 5.5.

Figura 5.5



Métodos de Runge-Kutta de orden 2

El primer paso para deducir un método Runge-Kutta es determinar los valores para a_1, α_1, y β_1 con la propiedad de que $a_1 f(t+\alpha_1, y+\beta_1)$ se aproxima a

$$T^{(2)}(t, y) = f(t, y) + \frac{h}{2} f'(t, y),$$

con error no mayor a $O(h^2)$, que es igual al orden del error de truncamiento local para el método de Taylor de orden 2. Ya que

$$f'(t, y) = \frac{df}{dt}(t, y) = \frac{\partial f}{\partial t}(t, y) + \frac{\partial f}{\partial y}(t, y) \cdot y'(t) \quad \text{y} \quad y'(t) = f(t, y),$$

tenemos

$$T^{(2)}(t, y) = f(t, y) + \frac{h}{2}\frac{\partial f}{\partial t}(t, y) + \frac{h}{2}\frac{\partial f}{\partial y}(t, y) \cdot f(t, y). \tag{5.18}$$

Al expandir $f(t+\alpha_1, y+\beta_1)$ en su polinomio de Taylor de grado 1, cerca de (t, y) obtenemos

$$\begin{aligned} a_1 f(t+\alpha_1, y+\beta_1) = a_1 f(t, y) + a_1\alpha_1 \frac{\partial f}{\partial t}(t, y) \\ + a_1\beta_1 \frac{\partial f}{\partial y}(t, y) + a_1 \cdot R_1(t+\alpha_1, y+\beta_1), \end{aligned} \tag{5.19}$$

donde

$$R_1(t+\alpha_1, y+\beta_1) = \frac{\alpha_1^2}{2}\frac{\partial^2 f}{\partial t^2}(\xi, \mu) + \alpha_1\beta_1 \frac{\partial^2 f}{\partial t \partial y}(\xi, \mu) + \frac{\beta_1^2}{2}\frac{\partial^2 f}{\partial y^2}(\xi, \mu), \tag{5.20}$$

para algunas ξ entre t y $t+\alpha_1$ y μ entre y y $y+\beta_1$.

Al ajustar los coeficientes de f y sus derivadas en las ecuaciones (5.18) y (5.19) obtenemos las tres ecuaciones

$$f(t, y): \ a_1 = 1; \qquad \frac{\partial f}{\partial t}(t, y): \ a_1\alpha_1 = \frac{h}{2}; \quad \text{y} \quad \frac{\partial f}{\partial y}(t, y): \ a_1\beta_1 = \frac{h}{2} f(t, y).$$

Los parámetros a_1, α_1, y β_1 son, por lo tanto

$$a_1 = 1, \quad \alpha_1 = \frac{h}{2}, \quad \text{y} \quad \beta_1 = \frac{h}{2} f(t, y),$$

por lo que

$$T^{(2)}(t, y) = f\left(t + \frac{h}{2}, y + \frac{h}{2} f(t, y)\right) - R_1\left(t + \frac{h}{2}, y + \frac{h}{2} f(t, y)\right),$$

y, a partir de la ecuación (5.20)

$$\begin{aligned} R_1\left(t + \frac{h}{2}, y + \frac{h}{2} f(t, y)\right) = \frac{h^2}{8}\frac{\partial^2 f}{\partial t^2}(\xi, \mu) + \frac{h^2}{4} f(t, y)\frac{\partial^2 f}{\partial t \partial y}(\xi, \mu) \\ + \frac{h^2}{8}(f(t, y))^2 \frac{\partial^2 f}{\partial y^2}(\xi, \mu). \end{aligned}$$

Si todas las derivadas parciales de segundo orden de f están acotadas, entonces

$$R_1\left(t+\frac{h}{2}, y+\frac{h}{2}f(t, y)\right)$$

es $O(h^2)$. En consecuencia:

- El orden de error para este nuevo método es igual al del método de Taylor de orden 2.

El método de ecuación de diferencia que resulta de reemplazar $T^{(2)}(t, y)$ en el método de Taylor de orden 2 por $f(t+(h/2), y+(h/2)f(t, y))$ es un método Runge-Kutta específico, conocido como *método de punto medio.*

Método de punto medio

$$w_0 = \alpha,$$

$$w_{i+1} = w_i + hf\left(t_i+\frac{h}{2}, w_i+\frac{h}{2}f(t_i, w_i)\right), \quad \text{para } i = 0, 1, \ldots, N-1.$$

Solamente se encuentran tres parámetros en $a_1 f(t+\alpha_1, y+\beta_1)$, y todos son necesarios para ajustar $T^{(2)}$. Por lo que se requiere una forma más complicada para satisfacer las condiciones para cualquiera de los métodos de Taylor de orden superior.

La forma de cuatro parámetros más adecuada para aproximar

$$T^{(3)}(t, y) = f(t, y) + \frac{h}{2}f'(t, y) + \frac{h^2}{6}f''(t, y)$$

es

$$a_1 f(t, y) + a_2 f(t+\alpha_2, y+\delta_2 f(t, y)), \tag{5.21}$$

e incluso con esto, no hay suficiente flexibilidad para ajustar el término

$$\frac{h^2}{6}\left[\frac{\partial f}{\partial y}(t, y)\right]^2 f(t, y),$$

lo cual resulta en la expansión de $(h^2/6)f''(t, y)$. Por consiguiente, lo mejor que se puede obtener al usar (5.21) son métodos con error de truncamiento local $O(h^2)$.

Sin embargo, el hecho de que (5.21) tenga cuatro parámetros proporciona una flexibilidad en su elección, por lo que se puede derivar una serie de métodos $O(h^2)$. Uno de los más importantes es el *método modificado de Euler*, que corresponde a seleccionar $a_1 = a_2 = \frac{1}{2}$ y $\alpha_2 = \delta_2 = h$. Éste tiene la siguiente forma de ecuación de diferencia.

Método modificado de Euler

$$w_0 = \alpha,$$

$$w_{i+1} = w_i + \frac{h}{2}[f(t_i, w_i) + f(t_{i+1}, w_i + hf(t_i, w_i))], \quad \text{para} \quad i = 0, 1, \ldots, N-1.$$

Ejemplo 2 Use los métodos de punto medio y modificado de Euler con $N = 10, h = 0.2, t_i = 0.2i$, y $w_0 = 0.5$ para aproximar la solución de nuestro ejemplo habitual

$$y' = y - t^2 + 1, \quad 0 \leq t \leq 2, \quad y(0) = 0.5.$$

Solución Las ecuaciones de diferencia producidas a partir de las diferentes fórmulas son

$$\text{método de punto medio: } w_{i+1} = 1.22w_i - 0.0088i^2 - 0.008i + 0.218$$

y

$$\text{método modificado de Euler: } w_{i+1} = 1.22w_i - 0.0088i^2 - 0.008i + 0.216,$$

Para cada $i = 0, 1, \ldots, 9$. Los primeros dos pasos de estos métodos nos dan

$$\text{método de punto medio: } w_1 = 1.22(0.5) - 0.0088(0)^2 - 0.008(0) + 0.216 = 0.826$$

y

$$\text{método modificado de Euler: } w_1 = 1.22(0.5) - 0.0088(0)^2 - 0.008(0) + 0.218 = 0.828$$

y

$$\text{método de punto medio: } w_2 = 1.22(0.828) - 0.0088(0.2)^2 - 0.008(0.2) + 0.218$$
$$= 1.21136$$

y

$$\text{método modificado de Euler: } w_2 = 1.22(0.826) - 0.0088(0.2)^2 - 0.008(0.2) + 0.216$$
$$= 1.20692.$$

La tabla 5.6 enumera todos los resultados de los cálculos. Para este problema, el método de punto medio es superior al método modificado de Euler. ■

Tabla 5.6

t_i	$y(t_i)$	Método de punto medio	Error	Método modificado de Euler	Error
0.0	0.5000000	0.5000000	0	0.5000000	0
0.2	0.8292986	0.8280000	0.0012986	0.8260000	0.0032986
0.4	1.2140877	1.2113600	0.0027277	1.2069200	0.0071677
0.6	1.6489406	1.6446592	0.0042814	1.6372424	0.0116982
0.8	2.1272295	2.1212842	0.0059453	2.1102357	0.0169938
1.0	2.6408591	2.6331668	0.0076923	2.6176876	0.0231715
1.2	3.1799415	3.1704634	0.0094781	3.1495789	0.0303627
1.4	3.7324000	3.7211654	0.0112346	3.6936862	0.0387138
1.6	4.2834838	4.2706218	0.0128620	4.2350972	0.0483866
1.8	4.8151763	4.8009586	0.0142177	4.7556185	0.0595577
2.0	5.3054720	5.2903695	0.0151025	5.2330546	0.0724173

Métodos de Runge-Kutta de orden superior

El término $T^{(3)}(t, y)$ se puede aproximar con error $O(h^3)$ mediante una expresión de la forma

$$f(t + \alpha_1, y + \delta_1 f(t + \alpha_2, y + \delta_2 f(t, y))),$$

relacionada con cuatro parámetros y el álgebra implicada en la determinación de $\alpha_1, \delta_1, \alpha_2,$ y δ_2 es bastante tediosa. El método $O(h^3)$ más común es el de Heun, dado por

Karl Heun (1859–1929) fue un profesor de la Technical University of Karlsruhe. Presentó esta técnica en un artículo publicado en 1900. [Heu]

$$w_0 = \alpha$$

$$w_{i+1} = w_i + \frac{h}{4}\left(f(t_i, w_i) + 3\left(f\left(t_i + \frac{2h}{3}, w_i + \frac{2h}{3} f\left(t_i + \frac{h}{3}, w_i + \frac{h}{3} f(t_i, w_i)\right)\right)\right)\right),$$
$$\text{para} \quad i = 0, 1, \ldots, N - 1.$$

Ilustración Al aplicar el método de Heun con $N = 10$, $h = 0.2$, $t_i = 0.2i$, y $w_0 = 0.5$ para aproximar la solución a nuestro ejemplo habitual

$$y' = y - t^2 + 1, \quad 0 \leq t \leq 2, \quad y(0) = 0.5,$$

da los valores en la tabla 5.7. Observe el error reducido a lo largo del rango sobre las aproximaciones de punto medio y Euler modificado. ■

Tabla 5.7

t_i	$y(t_i)$	Método de Heun	Error
0.0	0.5000000	0.5000000	0
0.2	0.8292986	0.8292444	0.0000542
0.4	1.2140877	1.2139750	0.0001127
0.6	1.6489406	1.6487659	0.0001747
0.8	2.1272295	2.1269905	0.0002390
1.0	2.6408591	2.6405555	0.0003035
1.2	3.1799415	3.1795763	0.0003653
1.4	3.7324000	3.7319803	0.0004197
1.6	4.2834838	4.2830230	0.0004608
1.8	4.8151763	4.8146966	0.0004797
2.0	5.3054720	5.3050072	0.0004648

En general, los métodos de Runge-Kutta de orden 3 no se usan. El método Runge-Kutta que se usa de manera común es de orden 4 en forma de ecuación de diferencia, dado como sigue.

Runge-Kutta de orden 4

$$w_0 = \alpha,$$
$$k_1 = hf(t_i, w_i),$$
$$k_2 = hf\left(t_i + \frac{h}{2}, w_i + \frac{1}{2}k_1\right),$$
$$k_3 = hf\left(t_i + \frac{h}{2}, w_i + \frac{1}{2}k_2\right),$$
$$k_4 = hf(t_{i+1}, w_i + k_3),$$
$$w_{i+1} = w_i + \frac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4),$$

para cada $i = 0, 1, \ldots, N-1$. Este método tiene error de truncamiento local $O(h^4)$, siempre y cuando la solución $y(t)$ tenga cinco derivadas continuas. Introducimos en el método la notación k_1, k_2, k_3, k_4 para eliminar la necesidad de anidado sucesivo en la segunda variable de $f(t, y)$. El ejercicio 32 muestra qué tan complicado se vuelve este anidado.

El algoritmo 5.2 implementa el método Runge-Kutta de orden 4.

ALGORITMO 5.2

Método Runge-Kutta (orden 4)

Para aproximar la solución del problema de valor inicial

$$y' = f(t, y), \quad a \leq t \leq b, \quad y(a) = \alpha,$$

en $(N + 1)$ números espaciados equitativamente en el intervalo$[a, b]$:

ENTRADA extremos a, b; entero N; condición inicial α.

SALIDA aproximación w para y en los valores $(N + 1)$ de t.

Paso 1 Determine $h = (b - a)/N$;
$t = a$;
$w = \alpha$;
SALIDA (t, w).

Paso 2 Para $i = 1, 2, \ldots, N$ haga los pasos 3–5.

Paso 3 Determine $K_1 = hf(t, w)$;
$K_2 = hf(t + h/2, w + K_1/2)$;
$K_3 = hf(t + h/2, w + K_2/2)$;
$K_4 = hf(t + h, w + K_3)$.

Paso 4 Determine $w = w + (K_1 + 2K_2 + 2K_3 + K_4)/6$; (*calcule* w_i.)
$t = a + ih$. (*Calcule* t_i.)

Paso 5 SALIDA (t, w).

Paso 6 PARE. ■

Ejemplo 3 Utilice el método Runge-Kutta de orden 4 con $h = 0.2$, $N = 10$, y $t_i = 0.2i$ para obtener aproximaciones para la solución del problema de valor inicial

$$y' = y - t^2 + 1, \quad 0 \le t \le 2, \quad y(0) = 0.5.$$

Solución La aproximación para $y(0.2)$ se obtiene mediante

$$w_0 = 0.5$$

$$k_1 = 0.2f(0, 0.5) = 0.2(1.5) = 0.3$$

$$k_2 = 0.2f(0.1, 0.65) = 0.328$$

$$k_3 = 0.2f(0.1, 0.664) = 0.3308$$

$$k_4 = 0.2f(0.2, 0.8308) = 0.35816$$

$$w_1 = 0.5 + \frac{1}{6}(0.3 + 2(0.328) + 2(0.3308) + 0.35816) = 0.8292933.$$

Los resultados restantes y sus errores se muestran en la tabla 5.8. ■

Tabla 5.8

t_i	Exacto $y_i = y(t_i)$	Runge-Kutta orden 4 w_i	Error $\lvert y_i - w_i \rvert$
0.0	0.5000000	0.5000000	0
0.2	0.8292986	0.8292933	0.0000053
0.4	1.2140877	1.2140762	0.0000114
0.6	1.6489406	1.6489220	0.0000186
0.8	2.1272295	2.1272027	0.0000269
1.0	2.6408591	2.6408227	0.0000364
1.2	3.1799415	3.1798942	0.0000474
1.4	3.7324000	3.7323401	0.0000599
1.6	4.2834838	4.2834095	0.0000743
1.8	4.8151763	4.8150857	0.0000906
2.0	5.3054720	5.3053630	0.0001089

Comparaciones computacionales

El principal esfuerzo computacional al aplicar los métodos de Runge-Kutta es la evaluación de f. En los métodos de segundo orden, el error de truncamiento local es $O(h^2)$, y el costo es dos evaluaciones de función por paso. El método Runge-Kutta de orden 4 requiere cuatro evaluaciones por paso, y el error de truncamiento local es $O(h^4)$. Butcher (consulte [But] para un resumen) ha establecido la relación entre el número de evaluaciones por paso y el orden del error de truncamiento local mostrado en la tabla 5.9. Esta tabla indica porqué los métodos de orden menor a cinco con tamaño de paso más pequeño se usan preferentemente para los métodos de orden superior por medio de un tamaño de paso más grande.

Tabla 5.9

Evaluaciones por paso	2	3	4	$5 \leq n \leq 7$	$8 \leq n \leq 9$	$10 \leq n$
Mejor error de truncamiento local posible	$O(h^2)$	$O(h^3)$	$O(h^4)$	$O(h^{n-1})$	$O(h^{n-2})$	$O(h^{n-3})$

Una medida para comparar los métodos de Runge-Kutta de orden inferior se describe como sigue:

- Mientras el método Runge-Kutta de orden 4 requiere cuatro evaluaciones por paso, el método de Euler sólo requiere una evaluación. Por lo tanto, si el método Runge-Kutta de orden 4 va a ser superior, deberían proporcionarse respuestas más precisas que las del método de Euler con un cuarto del tamaño de paso. De igual forma, si el método Runge-Kutta de orden 4 va a ser superior a los métodos de Runge-Kutta de segundo orden, que requieren dos evaluaciones por paso, debería proporcionar mayor precisión con un tamaño de paso h que un método de segundo orden con un tamaño de paso $h/2$.

Lo siguiente ilustra la superioridad del método Runge-Kutta de cuarto orden por medio de esta medida para el problema de valor inicial que hemos considerado.

Ilustración Para el problema

$$y' = y - t^2 + 1, \quad 0 \leq t \leq 2, \quad y(0) = 0.5,$$

el método de Euler con $h = 0.025$, el método de punto medio con $h = 0.05$, y el método Runge-Kutta de cuarto orden con $h = 0.1$ se comparan en los puntos de malla comunes de estos métodos 0.1, 0.2, 0.3, 0.4, y 0.5. Cada una de estas técnicas requiere 20 evaluaciones de función para determinar los valores enumerados en la tabla 5.10 para aproximar $y(0.5)$. En este ejemplo, el método de cuarto orden es claramente superior. ■

Tabla 5.10

t_i	Exacto	Euler $h = 0.025$	Euler modificado $h = 0.05$	Runge-Kutta de orden 4 $h = 0.1$
0.0	0.5000000	0.5000000	0.5000000	0.5000000
0.1	0.6574145	0.6554982	0.6573085	0.6574144
0.2	0.8292986	0.8253385	0.8290778	0.8292983
0.3	1.0150706	1.0089334	1.0147254	1.0150701
0.4	1.2140877	1.2056345	1.2136079	1.2140869
0.5	1.4256394	1.4147264	1.4250141	1.4256384

La sección Conjunto de ejercicios 5.4 está disponible en línea. Encuentre la ruta de acceso en las páginas preliminares.



5.5 Control de error y método Runge-Kutta-Fehlberg

Tal vez le gustaría revisar el material sobre cuadratura adaptable en la sección 4.6 antes de considerar este material.

En la sección 4.6 observamos el uso apropiado de tamaños de paso variables para que las aproximaciones de integrales produzcan métodos eficientes. Por sí mismos quizá no fueran suficientes para favorecer estos métodos debido al aumento de complicación que surge al aplicarlos. Sin embargo, tienen otra característica que les da gran valor. En el procedimiento de tamaño de paso se incluye un cálculo del error de truncamiento que no requiere la aproximación de derivadas superiores de la función. Estos métodos reciben el nombre de *adaptables* porque adaptan el número y la posición de los nodos utilizados en la aproximación para garantizar que el error de truncamiento se mantiene dentro de un límite específico.

Existe una conexión cercana entre el problema de aproximación del valor de una integral definida y el de aproximación de la solución de un problema de valor inicial. No es sorprendente, entonces, que existan métodos adaptables para aproximar las soluciones de los problemas de valor inicial y que estos métodos no sólo sean eficientes, sino también que incluyan el control de error.

Cualquier método de un paso para aproximar la solución, $y(t)$, del problema de valor inicial

$$y' = f(t, y), \quad \text{para } a \leq t \leq b, \quad \text{con } y(a) = \alpha,$$

se puede expresar en la forma

$$w_{i+1} = w_i + h_i \phi(t_i, w_i, h_i), \quad \text{para } i = 0, 1, \ldots, N-1,$$

para alguna función ϕ.

Un método ideal para la ecuación de diferencia

$$w_{i+1} = w_i + h_i \phi(t_i, w_i, h_i), \quad i = 0, 1, \ldots, N-1,$$

para aproximar la solución, $y(t)$, del problema de valor inicial

$$y' = f(t, y), \quad a \leq t \leq b, \quad y(a) = \alpha,$$

tendría la propiedad de que, dada una tolerancia $\varepsilon > 0$, un número mínimo de puntos de malla se puede usar para garantizar que el error global, $|y(t_i) - w_i|$, no exceda ε para ninguna $i = 0, 1, \ldots, N$. No sorprende que tener un número mínimo de puntos de malla y también controlar el error global de un método de diferencia sea inconsistente con los puntos igualmente espaciados en el intervalo. En esta sección examinamos técnicas que se usan para controlar el error de un método de ecuación de diferencia de manera eficiente mediante la elección adecuada de puntos de malla.

A pesar de que, en general, no podamos determinar el error global de un método, en la sección 5.10 observaremos que existe una conexión cercana entre el error de truncamiento local y el error global. Al utilizar métodos de diferente orden podemos predecir el error de truncamiento local y, al usar esta predicción, seleccionar un tamaño de paso que controle al error global.

Para ilustrar la técnica, suponga que tenemos dos técnicas de aproximación. La primera se obtiene a partir del método de Taylor de enésimo orden de la forma

$$y(t_{i+1}) = y(t_i) + h\phi(t_i, y(t_i), h) + O(h^{n+1})$$

y produce aproximaciones con error de truncamiento local $\tau_{i+1}(h) = O(h^n)$. Éstas son dadas por

$$w_0 = \alpha$$
$$w_{i+1} = w_i + h\phi(t_i, w_i, h), \quad \text{para } i > 0.$$

En general, el método se genera al aplicar la modificación de Runge-Kutta para el método de Taylor, pero la derivación específica no es importante.

El segundo método es similar, pero de un orden superior; proviene de un método de Taylor de $(n + 1)$ orden de la forma

$$y(t_{i+1}) = y(t_i) + h\tilde{\phi}(t_i, y(t_i), h) + O(h^{n+2})$$

y produce aproximaciones con el error de truncamiento local $\tilde{\tau}_{i+1}(h) = O(h^{n+1})$. Esto es dado por

$$\tilde{w}_0 = \alpha$$
$$\tilde{w}_{i+1} = \tilde{w}_i + h\tilde{\phi}(t_i, \tilde{w}_i, h), \quad \text{para } i > 0.$$

Primero suponemos que $w_i \approx y(t_i) \approx \tilde{w}_i$ y seleccionamos un tamaño de paso fijo h para generar las aproximaciones w_{i+1} y $\tilde{w}_{i+1}$ to $y(t_{i+1})$. Entonces

$$\begin{aligned}
\tau_{i+1}(h) &= \frac{y(t_{i+1}) - y(t_i)}{h} - \phi(t_i, y(t_i), h) \\
&= \frac{y(t_{i+1}) - w_i}{h} - \phi(t_i, w_i, h) \\
&= \frac{y(t_{i+1}) - [w_i + h\phi(t_i, w_i, h)]}{h} \\
&= \frac{1}{h}(y(t_{i+1}) - w_{i+1}).
\end{aligned}$$

De manera similar, tenemos

$$\tilde{\tau}_{i+1}(h) = \frac{1}{h}(y(t_{i+1}) - \tilde{w}_{i+1}).$$

Como consecuencia, tenemos

$$\begin{aligned}
\tau_{i+1}(h) &= \frac{1}{h}(y(t_{i+1}) - w_{i+1}) \\
&= \frac{1}{h}[(y(t_{i+1}) - \tilde{w}_{i+1}) + (\tilde{w}_{i+1} - w_{i+1})] \\
&= \tilde{\tau}_{i+1}(h) + \frac{1}{h}(\tilde{w}_{i+1} - w_{i+1}).
\end{aligned}$$

Pero $\tau_{i+1}(h)$ es $O(h^n)$ y $\tilde{\tau}_{i+1}(h)$ es $O(h^{n+1})$, por lo que la parte significativa de $\tau_{i+1}(h)$ debe provenir de

$$\frac{1}{h}\left(\tilde{w}_{i+1} - w_{i+1}\right).$$

Esto nos da una aproximación que se puede calcular fácilmente para el método de error de truncamiento local de $O(h^n)$:

$$\tau_{i+1}(h) \approx \frac{1}{h}\left(\tilde{w}_{i+1} - w_{i+1}\right).$$

Sea $R = \frac{1}{h}|\tilde{w}_{i+1} - w_{i+1}|$.

Sin embargo, el objetivo no es simplemente calcular el error de truncamiento local, sino ajustar el tamaño de paso para mantenerlo dentro de una cota específica. Para hacerlo, asumimos que como $\tau_{i+1}(h)$ es $O(h^n)$, existe un número K, independiente de h, con

$$\tau_{i+1}(h) \approx Kh^n.$$

Entonces, el error de truncamiento local producido al aplicar el método de enésimo orden con un tamaño de paso qh se puede calcular a través de las aproximaciones originales w_{i+1} y $\tilde{w}_{i+1}$:

$$\tau_{i+1}(qh) \approx K(qh)^n = q^n(Kh^n) \approx q^n \tau_{i+1}(h) \approx \frac{q^n}{h}(\tilde{w}_{i+1} - w_{i+1}).$$

La cota $\tau_{i+1}(qh)$ para ε, seleccionamos q de tal forma que

$$\frac{q^n}{h}|\tilde{w}_{i+1} - w_{i+1}| \approx |\tau_{i+1}(qh)| \leq \varepsilon,$$

es decir, tal que

$$q \leq \left(\frac{\varepsilon h}{|\tilde{w}_{i+1} - w_{i+1}|}\right)^{1/n} = \left(\frac{\epsilon}{R}\right)^{1/n}. \tag{5.22}$$

Método Runge-Kutta-Fehlberg

Erwin Fehlberg desarrolló ésta y otras técnicas de control de error mientras trabajaba en las instalaciones de la NASA en Huntsville, Alabama, durante la década de 1960. En 1969, recibió por su trabajo la medalla por logros científicos excepcionales de la NASA.

Una técnica popular que usa la desigualdad (5.22) para control de error es el **método Runge-Kutta-Fehlberg** (consulte [Fe].) Esta técnica utiliza el método Runge-Kutta con error de truncamiento local de orden 5,

$$\tilde{w}_{i+1} = w_i + \frac{16}{135}k_1 + \frac{6656}{12825}k_3 + \frac{28561}{56430}k_4 - \frac{9}{50}k_5 + \frac{2}{55}k_6,$$

para calcular el error local en un método Runge-Kutta de orden 4 dado por

$$w_{i+1} = w_i + \frac{25}{216}k_1 + \frac{1408}{2565}k_3 + \frac{2197}{4104}k_4 - \frac{1}{5}k_5,$$

donde los coeficientes de las ecuaciones son

$$k_1 = hf(t_i, w_i),$$

$$k_2 = hf\left(t_i + \frac{h}{4}, w_i + \frac{1}{4}k_1\right),$$

$$k_3 = hf\left(t_i + \frac{3h}{8}, w_i + \frac{3}{32}k_1 + \frac{9}{32}k_2\right),$$

$$k_4 = hf\left(t_i + \frac{12h}{13}, w_i + \frac{1932}{2197}k_1 - \frac{7200}{2197}k_2 + \frac{7296}{2197}k_3\right),$$

$$k_5 = hf\left(t_i + h, w_i + \frac{439}{216}k_1 - 8k_2 + \frac{3680}{513}k_3 - \frac{845}{4104}k_4\right),$$

$$k_6 = hf\left(t_i + \frac{h}{2}, w_i - \frac{8}{27}k_1 + 2k_2 - \frac{3544}{2565}k_3 + \frac{1859}{4104}k_4 - \frac{11}{40}k_5\right).$$

Una ventaja de este método es que sólo se requieren seis evaluaciones de f por paso. Los métodos arbitrarios de Runge-Kutta de órdenes 4 y 5, utilizados juntos (consulte la tabla 5.9 en la página 217) requieren por lo menos cuatro evaluaciones de f para el método de cuarto orden y un sexto adicional para el método de quinto orden, para un total de por lo menos 10 evaluaciones de función. Por lo que el método Runge-Kutta-Fehlberg tiene al menos 40% de disminución en el número de evaluaciones de función sobre el uso de un par de métodos arbitrarios de cuarto y quinto orden.

En la teoría de control de error, un valor inicial de h en el i-ésimo paso se usa para encontrar los primeros valores de w_{i+1} y $\tilde{w}_{i+1}$, lo cual conduce a la determinación de q para ese paso y, entonces, se repiten los cálculos. Este procedimiento requiere el doble de evaluaciones de función por paso sin control de error. En la práctica, el valor de q que se utilizará

se selecciona de una forma diferente con el fin de hacer que el costo de la evaluación de función incrementado valga la pena. El valor de q determinado en el i-ésimo paso se utiliza para dos propósitos:

- Cuando $R > \epsilon$, rechazamos la selección inicial de h en el i-ésimo paso y repetimos los cálculos mediante qh y
- Cuando $R \leq \epsilon$, aceptamos el valor calculado en el i-ésimo paso mediante el tamaño de paso h, pero cambiamos el tamaño de paso para qh para el $(i + 1)$ paso.

Debido a la desventaja en términos de evaluaciones de función que se debe pagar si los pasos se repiten, q tiende a ser seleccionado de manera conservadora. De hecho, para el método Runge-Kutta-Fehlberg con $n = 4$, una elección común es

$$q = \left(\frac{\varepsilon h}{2|\tilde{w}_{i+1} - w_{i+1}|}\right)^{1/4} = 0.84\left(\frac{\varepsilon h}{|\tilde{w}_{i+1} - w_{i+1}|}\right)^{1/4} = 0.84\left(\frac{\epsilon}{R}\right)^{1/n}.$$

En el algoritmo 5.3 para el método Runge-Kutta-Fehlberg, se añade el paso 9 para eliminar grandes modificaciones en el tamaño de paso. Esto se hace para no pasar demasiado tiempo con tamaños de paso pequeños en regiones con irregularidades en las derivadas de y y para evitar tamaños de paso grandes, lo cual puede resultar en la omisión de regiones sensibles entre los pasos. El procedimiento de incremento del tamaño de paso se puede evitar por completo a partir del algoritmo y el procedimiento de disminución del tamaño de paso que se usa sólo cuando sea necesario para controlar el error.

ALGORITMO **5.3**

Método Runge-Kutta-Fehlberg

Para aproximar la solución del problema de valor inicial

$$y' = f(t, y), \quad a \leq t \leq b, \quad y(a) = \alpha,$$

con error de truncamiento local dentro de una tolerancia determinada:

ENTRADA extremos a, b; condición inicial α; tolerancia TOL; tamaño de paso máximo $hmáx$; tamaño de paso mínimo $hmín$.

SALIDA t, w, h, donde w se aproxima a $y(t)$ y se utiliza el tamaño de paso h o un mensaje de que se excede el tamaño mínimo de paso.

Paso 1 Determine $t = a$;
$w = \alpha$;
$h = hmáx$;
$FLAG = 1$;
SALIDA (t, w).

Paso 2 Mientras $(FLAG = 1)$ haga los pasos 3–11.

Paso 3 Determine $K_1 = hf(t, w)$;

$$K_2 = hf\left(t + \tfrac{1}{4}h, w + \tfrac{1}{4}K_1\right);$$
$$K_3 = hf\left(t + \tfrac{3}{8}h, w + \tfrac{3}{32}K_1 + \tfrac{9}{32}K_2\right);$$
$$K_4 = hf\left(t + \tfrac{12}{13}h, w + \tfrac{1932}{2197}K_1 - \tfrac{7200}{2197}K_2 + \tfrac{7296}{2197}K_3\right);$$
$$K_5 = hf\left(t + h, w + \tfrac{439}{216}K_1 - 8K_2 + \tfrac{3680}{513}K_3 - \tfrac{845}{4104}K_4\right);$$
$$K_6 = hf\left(t + \tfrac{1}{2}h, w - \tfrac{8}{27}K_1 + 2K_2 - \tfrac{3544}{2565}K_3 + \tfrac{1859}{4104}K_4 - \tfrac{11}{40}K_5\right).$$

Paso 4 Determine $R = \frac{1}{h}|\frac{1}{360}K_1 - \frac{128}{4275}K_3 - \frac{2197}{75240}K_4 + \frac{1}{50}K_5 + \frac{2}{55}K_6|$.

(*Nota:* $R = \frac{1}{h}|\tilde{w}_{i+1} - w_{i+1}| \approx |\tau_{i+1}(h)|$.)

Paso 5 Si $R \leq TOL$ entonces haga los pasos 6 y 7.

Paso 6 Determine $t = t + h$; (*aproximación aceptada*.)

$$w = w + \frac{25}{216}K_1 + \frac{1408}{2565}K_3 + \frac{2197}{4104}K_4 - \frac{1}{5}K_5.$$

Paso 7 SALIDA (t, w, h). (*Fin del paso 5*)

Paso 8 Determine $\delta = 0.84(TOL/R)^{1/4}$.

Paso 9 Si $\delta \leq 0.1$ entonces haga $h = 0.1h$
también si $\delta \geq 4$ entonces haga $h = 4h$
si no haga $h = \delta h$. (*Calcular h nueva*.)

Paso 10 Si $h > hmáx$ entonces haga $h = hmáx$.

Paso 11 Si $t \geq b$ entonces haga $FLAG = 0$
también si $t + h > b$ entonces haga $h = b - t$
también si $h < hmín$ entonces
determine $FLAG = 0$;
SALIDA (‘*h mínima excedida*’).
(*Procedimiento completado sin éxito*.)
(*Fin del paso 3*)

Paso 12 (*El procedimiento está completo*.)
PARE. ▪

Ejemplo 1 Utilice el método Runge-Kutta-Fehlberg con una tolerancia $TOL = 10^{-5}$, un tamaño de paso máximo $hmáx = 0.25$, un tamaño de paso mínimo $hmín = 0.01$ para aproximar la solución del problema de valor inicial

$$y' = y - t^2 + 1, \quad 0 \leq t \leq 2, \quad y(0) = 0.5,$$

y compare los resultados con la solución exacta $y(t) = (t + 1)^2 - 0.5e^t$.

Solución Trabajaremos a través del primer paso de los cálculos y, a continuación, aplicaremos el algoritmo 5.3 para determinar los resultados restantes. La condición inicial da $t_0 = 0$ y $w_0 = 0.5$. Para determinar w_1 mediante w_1 usando $h = 0.25$, el tamaño de paso máximo permisible, calculamos

$$k_1 = hf(t_0, w_0) = 0.25\left(0.5 - 0^2 + 1\right) = 0.375,$$

$$k_2 = hf\left(t_0 + \frac{1}{4}h, w_0 + \frac{1}{4}k_1\right) = 0.25f\left(\frac{1}{4}0.25, 0.5 + \frac{1}{4}0.375\right) = 0.3974609,$$

$$k_3 = hf\left(t_0 + \frac{3}{8}h, w_0 + \frac{3}{32}k_1 + \frac{9}{32}k_2\right)$$

$$= 0.25f\left(0.09375, 0.5 + \frac{3}{32}0.375 + \frac{9}{32}0.3974609\right) = 0.4095383,$$

$$k_4 = hf\left(t_0 + \frac{12}{13}h, w_0 + \frac{1932}{2197}k_1 - \frac{7200}{2197}k_2 + \frac{7296}{2197}k_3\right)$$

$$= 0.25f\left(0.2307692, 0.5 + \frac{1932}{2197}0.375 - \frac{7200}{2197}0.3974609 + \frac{7296}{2197}0.4095383\right)$$

$$= 0.4584971,$$

$$k_5 = hf\left(t_0 + h, w_0 + \frac{439}{216}k_1 - 8k_2 + \frac{3680}{513}k_3 - \frac{845}{4104}k_4\right)$$
$$= 0.25\,f\left(0.25, 0.5 + \frac{439}{216}0.375 - 8(0.3974609) + \frac{3680}{513}0.4095383 - \frac{845}{4104}0.4584971\right)$$
$$= 0.4658452,$$

y

$$k_6 = hf\left(t_0 + \frac{1}{2}h, w_0 - \frac{8}{27}k_1 + 2k_2 - \frac{3544}{2565}k_3 + \frac{1859}{4104}k_4 - \frac{11}{40}k_5\right)$$
$$= 0.25\,f\left(0.125, 0.5 - \frac{8}{27}0.375 + 2(0.3974609) - \frac{3544}{2565}0.4095383 + \frac{1859}{4104}0.4584971\right.$$
$$\left. - \frac{11}{40}0.4658452\right)$$
$$= 0.4204789.$$

Entonces, se encuentra que las dos aproximaciones para $y(0.25)$ son

$$\tilde{w}_1 = w_0 + \frac{16}{135}k_1 + \frac{6656}{12825}k_3 + \frac{28561}{56430}k_4 - \frac{9}{50}k_5 + \frac{2}{55}k_6$$
$$= 0.5 + \frac{16}{135}0.375 + \frac{6656}{12825}0.4095383 + \frac{28561}{56430}0.4584971 - \frac{9}{50}0.4658452$$
$$+ \frac{2}{55}0.4204789$$
$$= 0.9204870,$$

y

$$w_1 = w_0 + \frac{25}{216}k_1 + \frac{1408}{2565}k_3 + \frac{2197}{4104}k_4 - \frac{1}{5}k_5$$
$$= 0.5 + \frac{25}{216}0.375 + \frac{1408}{2565}0.4095383 + \frac{2197}{4104}0.4584971 - \frac{1}{5}0.4658452$$
$$= 0.9204886.$$

Esto también implica que

$$R = \frac{1}{0.25}\left|\frac{1}{360}k_1 - \frac{128}{4275}k_3 - \frac{2197}{75240}k_4 + \frac{1}{50}k_5 + \frac{2}{55}k_6\right|$$
$$= 4\left|\frac{1}{360}0.375 - \frac{128}{4275}0.4095383 - \frac{2197}{75240}0.4584971\right.$$
$$\left. + \frac{1}{50}0.4658452 + \frac{2}{55}0.4204789\right|$$
$$= 0.00000621388,$$

y

$$q = 0.84\left(\frac{\varepsilon}{R}\right)^{1/4} = 0.84\left(\frac{0.00001}{0.00000621388}\right)^{1/4} = 0.9461033291.$$

Puesto que $R \leq 10^{-5}$, podemos aceptar la aproximación 0.9204886 para $y(0.25)$, pero deberíamos ajustar el tamaño de paso para la siguiente iteración para $h = 0.9461033291$ $(0.25) \approx 0.2365258$. Sin embargo, sólo se esperaría que los primeros cinco dígitos de este resultado sean precisos porque R tiene solamente cinco dígitos de precisión. Puesto que estamos restando efectivamente los números casi iguales w_i y $\tilde{w}_i$ cuando calculamos R, existe una buena probabilidad de error de redondeo. Ésta es una razón adicional para ser conservador al calcular q.

Los resultados a partir del algoritmo se muestran en la tabla 5.11. El incremento de precisión se ha usado para garantizar que los cálculos son precisos para todos los lugares listados. Las últimas dos columnas en la tabla 5.11 muestran los resultados del método de quinto orden. Para valores pequeños de t, el error es menor al error en el método de cuarto orden, pero supera el del método de cuarto orden cuando t aumenta. ■

Tabla 5.11

t_i	$y_i = y(t_i)$	RKF-4 w_i	h_i	R_i	$\lvert y_i - w_i \rvert$	RKF-5 $\hat{w}_i$	$\lvert y_i - \hat{w}_i \rvert$
0	0.5	0.5			0.5		
0.2500000	0.9204873	0.9204886	0.2500000	6.2×10^{-6}	1.3×10^{-6}	0.9204870	2.424×10^{-7}
0.4865522	1.3964884	1.3964910	0.2365522	4.5×10^{-6}	2.6×10^{-6}	1.3964900	1.510×10^{-6}
0.7293332	1.9537446	1.9537488	0.2427810	4.3×10^{-6}	4.2×10^{-6}	1.9537477	3.136×10^{-6}
0.9793332	2.5864198	2.5864260	0.2500000	3.8×10^{-6}	6.2×10^{-6}	2.5864251	5.242×10^{-6}
1.2293332	3.2604520	3.2604605	0.2500000	2.4×10^{-6}	8.5×10^{-6}	3.2604599	7.895×10^{-6}
1.4793332	3.9520844	3.9520955	0.2500000	7×10^{-7}	1.11×10^{-5}	3.9520954	1.096×10^{-5}
1.7293332	4.6308127	4.6308268	0.2500000	1.5×10^{-6}	1.41×10^{-5}	4.6308272	1.446×10^{-5}
1.9793332	5.2574687	5.2574861	0.2500000	4.3×10^{-6}	1.73×10^{-5}	5.2574871	1.839×10^{-5}
2.0000000	5.3054720	5.3054896	0.0206668		1.77×10^{-5}	5.3054896	1.768×10^{-5}

La sección Conjunto de ejercicios 5.5 está disponible en línea. Encuentre la ruta de acceso en las páginas preliminares.

5.6 Métodos multipasos

Los métodos que se han analizado hasta este punto en el capítulo reciben el nombre de **métodos de un paso** porque la aproximación para el punto de malla t_{i+1} involucra información de un solo punto de malla previo t_i. A pesar de que estos métodos podrían utilizar la información de evaluación de función en los puntos entre t_i y t_{i+1}, no la retienen para uso directo en aproximaciones futuras. Toda la información que se ha usado con estos métodos se obtiene dentro del subintervalo sobre el que la solución se aproxima.

La solución aproximada está disponible en cada uno de los puntos de malla $t_0, t_1, \ldots, t_i$ antes de obtener la aproximación en t_{i+1}, y porque el error $|w_j - y(t_j)|$ tiende a incrementar con j, por lo que parece razonable desarrollar métodos que usen datos previos más precisos al aproximar la solución t_{i+1}.

Los métodos que utilizan la aproximación en más de un punto de malla previo para determinar la aproximación en el siguiente punto reciben el nombre de métodos *multipasos*. A continuación se presenta la definición precisa de estos métodos, junto con la definición de los dos tipos de métodos multipasos.

Definición 5.14 Un **método multipasos de paso *m*** para resolver el problema de valor inicial

$$y' = f(t, y), \quad a \leq t \leq b, \quad y(a) = \alpha, \tag{5.23}$$

tiene una ecuación de diferencia para encontrar la aproximación w_{i+1} en el punto de malla t_{i+1} representado por la siguiente ecuación, donde m es un entero mayor que 1:

$$\begin{aligned} w_{i+1} = {} & a_{m-1}w_i + a_{m-2}w_{i-1} + \cdots + a_0 w_{i+1-m} \\ & + h[b_m f(t_{i+1}, w_{i+1}) + b_{m-1} f(t_i, w_i) \\ & + \cdots + b_0 f(t_{i+1-m}, w_{i+1-m})], \end{aligned} \tag{5.24}$$

para $i = m-1, m, \ldots, N-1$, donde $h = (b-a)/N$, $a_0, a_1, \ldots, a_{m-1}$ y $b_0, b_1, \ldots, b_m$ son constantes y se especifican los valores iniciales específicos

$$w_0 = \alpha, \quad w_1 = \alpha_1, \quad w_2 = \alpha_2, \quad \ldots, \quad w_{m-1} = \alpha_{m-1}$$

Son especificados ■

Cuando $b_m = 0$, el método recibe el nombre de **explícito**, o **abierto**, ya que la ecuación (5.24) proporciona w_{i+1} de manera explícita en términos de valores previamente determinados. Cuando $b_m \neq 0$, , el método recibe el nombre de **implícito**, o **cerrado**, porque w_{i+1} se presenta en ambos lados de la ecuación (5.24), por lo que w_{i+1} solamente se especifica de manera implícita.

Por ejemplo, las ecuaciones

Las técnicas de Adams-Bashforth se deben a John Couch Adams (1819–1892), quien realizó trabajos significativos en matemáticas y astronomía. Desarrolló estas técnicas numéricas para aproximar la solución de un problema de flujo de fluido propuesto por Bashforth.

$$w_0 = \alpha, \quad w_1 = \alpha_1, \quad w_2 = \alpha_2, \quad w_3 = \alpha_3,$$

$$w_{i+1} = w_i + \frac{h}{24}[55f(t_i, w_i) - 59f(t_{i-1}, w_{i-1}) + 37f(t_{i-2}, w_{i-2}) - 9f(t_{i-3}, w_{i-3})], \tag{5.25}$$

para cada $i = 3, 4, \ldots, N-1$, definen el método *explícito* de cuatro pasos conocido como **técnica Adams-Bashforth de cuarto orden**. Las ecuaciones

$$w_0 = \alpha, \quad w_1 = \alpha_1, \quad w_2 = \alpha_2,$$

$$w_{i+1} = w_i + \frac{h}{24}[9f(t_{i+1}, w_{i+1}) + 19f(t_i, w_i) - 5f(t_{i-1}, w_{i-1}) + f(t_{i-2}, w_{i-2})], \tag{5.26}$$

Forest Ray Moulton (1872–1952) estaba a cargo de balística en Aberdeen Proving Grounds en Maryland durante la Primera Guerra Mundial. Era un autor prolífico, escribió numerosos libros sobre matemáticas y astronomía y desarrolló métodos multipasos para resolver ecuaciones balísticas.

para cada $i = 2, 3, \ldots, N - 1$, definen un método *implícito* de tres pasos conocido como la **técnica de Adams-Moulton de cuarto orden**.

Los valores iniciales ya sea en la ecuación (5.25) como en la ecuación (5.26) deben especificarse, en general, al suponer $w_0 = \alpha$ y generar los valores restantes ya sea con el método Runge-Kutta o el método de Taylor. Observaremos que, en general, los métodos implícitos son más precisos que los explícitos, pero para aplicar directamente uno implícito como el (5.25), debemos resolver la ecuación implícita para w_{i+1}. Esto no siempre es posible e incluso cuando se puede hacer, la solución para w_{i+1} puede no ser única.

Ejemplo 1 En el ejemplo 3 de la sección 5.4 (consulte la tabla 5.8 en la página 216), usamos el método Runge-Kutta de orden 4 con $h = 0.2$ para aproximar soluciones para el problema de valor inicial

$$y' = y - t^2 + 1, \quad 0 \leq t \leq 2, \quad y(0) = 0.5.$$

Se encontró que las primeras cuatro aproximaciones son $y(0) = w_0 = 0.5$, $y(0.2) \approx w_1 = 0.8292933$, $y(0.4) \approx w_2 = 1.2140762$, y $y(0.6) \approx w_3 = 1.6489220$. Úselas como valores iniciales para el método Adam-Bashforth de cuarto orden para calcular nuevas aproximaciones para $y(0.8)$ y $y(1.0)$ y compárelas con las producidas mediante el método Runge-Kutta de cuarto orden.

Solución Para el método Adam-Bashforth de cuarto orden, tenemos

$$
\begin{aligned}
y(0.8) \approx w_4 = w_3 + \frac{0.2}{24}(55f(0.6, w_3) - 59f(0.4, w_2) + 37f(0.2, w_1) - 9f(0, w_0))\\
&= 1.6489220 + \frac{0.2}{24}(55f(0.6, 1.6489220) - 59f(0.4, 1.2140762)\\
&\quad + 37f(0.2, 0.8292933) - 9f(0, 0.5))\\
&= 1.6489220 + 0.0083333(55(2.2889220) - 59(2.0540762)\\
&\quad + 37(1.7892933) - 9(1.5))\\
&= 2.1272892
\end{aligned}
$$

y

$$
\begin{aligned}
y(1.0) \approx w_5 = w_4 + \frac{0.2}{24}(55f(0.8, w_4) - 59f(0.6, w_3) + 37f(0.4, w_2) - 9f(0.2, w_1))\\
&= 2.1272892 + \frac{0.2}{24}(55f(0.8, 2.1272892) - 59f(0.6, 1.6489220)\\
&\quad + 37f(0.4, 1.2140762) - 9f(0.2, 0.8292933))\\
&= 2.1272892 + 0.0083333(55(2.4872892) - 59(2.2889220)\\
&\quad + 37(2.0540762) - 9(1.7892933))\\
&= 2.6410533.
\end{aligned}
$$

El error para estas aproximaciones en $t = 0.8$ y $t = 1.0$ son, respectivamente,

$$|2.1272295 - 2.1272892| = 5.97 \times 10^{-5} \text{ y } |2.6410533 - 2.6408591| = 1.94 \times 10^{-4}.$$

Las aproximaciones correspondientes tienen los errores

$$|2.1272027 - 2.1272892| = 2.69 \times 10^{-5} \text{ y } |2.6408227 - 2.6408591| = 3.64 \times 10^{-5}.$$

■

Adams estaba especialmente interesado en usar esta habilidad para los cálculos numéricos precisos en la investigación de las órbitas de los planetas. Predijo la existencia de Neptuno al analizar las irregularidades en el planeta Urano y desarrolló numerosas técnicas de integración numérica para aproximar la solución de las ecuaciones diferenciales.

Para comenzar la deducción de un método multipasos observe que la solución para el problema de valor inicial

$$y' = f(t, y), \quad a \leq t \leq b, \quad y(a) = \alpha,$$

si se integra sobre el intervalo $[t_i, t_{i+1}]$, tiene la propiedad de que

$$y(t_{i+1}) - y(t_i) = \int_{t_i}^{t_{i+1}} y'(t)\,dt = \int_{t_i}^{t_{i+1}} f(t, y(t))\,dt.$$

Por consiguiente,

$$y(t_{i+1}) = y(t_i) + \int_{t_i}^{t_{i+1}} f(t, y(t))\,dt. \tag{5.27}$$

Sin embargo, no podemos integrar $f(t, y(t))$ sin conocer $y(t)$, la solución del problema, por lo que, en su lugar, integramos un polinomio interpolante $P(t)$ para $f(t, y(t))$, uno que está determinado por algunos de los puntos de datos previamente obtenidos $(t_0, w_0), (t_1, w_1), \ldots, (t_i, w_i)$. Cuando suponemos, además, que $y(t_i) \approx w_i$ la ecuación (5.27) se convierte en

$$y(t_{i+1}) \approx w_i + \int_{t_i}^{t_{i+1}} P(t)\, dt. \tag{5.28}$$

A pesar de que ninguna forma de polinomio de interpolación se puede usar para la derivación, es más conveniente utilizar la fórmula de diferencia regresiva porque esta forma incorpora con mayor facilidad los datos calculados más recientemente.

Para derivar una técnica de m pasos de Adam-Bashforth, formamos el polinomio de diferencia regresiva $P_{m-1}(t)$ con

$$(t_i, f(t_i, y(t_i))), \quad (t_{i-1}, f(t_{i-1}, y(t_{i-1}))), \ldots, \quad (t_{i+1-m}, f(t_{i+1-m}, y(t_{i+1-m}))).$$

Puesto que $P_{m-1}(t)$ es un polinomio de interpolación de grado $m-1$, existe algún número ξ_i en (t_{i+1-m}, t_i) con

$$f(t, y(t)) = P_{m-1}(t) + \frac{f^{(m)}(\xi_i, y(\xi_i))}{m!}(t-t_i)(t-t_{i-1})\cdots(t-t_{i+1-m}).$$

Al introducir la sustitución de la variable, $t = t_i + sh$, con $dt = h\, ds$, en $P_{m-1}(t)$ y el término de error implica que

$$\begin{aligned}\int_{t_i}^{t_{i+1}} f(t, y(t))\, dt &= \int_{t_i}^{t_{i+1}} \sum_{k=0}^{m-1} (-1)^k \binom{-s}{k} \nabla^k f(t_i, y(t_i))\, dt \\ &\quad + \int_{t_i}^{t_{i+1}} \frac{f^{(m)}(\xi_i, y(\xi_i))}{m!}(t-t_i)(t-t_{i-1})\cdots(t-t_{i+1-m})\, dt \\ &= \sum_{k=0}^{m-1} \nabla^k f(t_i, y(t_i)) h (-1)^k \int_0^1 \binom{-s}{k} ds \\ &\quad + \frac{h^{m+1}}{m!} \int_0^1 s(s+1)\cdots(s+m-1) f^{(m)}(\xi_i, y(\xi_i))\, ds.\end{aligned}$$

Tabla 5.12

k	$\int_0^1 (-1)^k \binom{-s}{k} ds$
0	1
1	$\frac{1}{2}$
2	$\frac{5}{12}$
3	$\frac{3}{8}$
4	$\frac{251}{720}$
5	$\frac{95}{288}$

Las integrales $(-1)^k \int_0^1 \binom{-s}{k}$ para diferentes valores de k se evalúan fácilmente y son listadas en la tabla 5.12. Por ejemplo, cuando $k = 3$,

$$\begin{aligned}(-1)^3 \int_0^1 \binom{-s}{3} ds &= -\int_0^1 \frac{(-s)(-s-1)(-s-2)}{1 \cdot 2 \cdot 3} ds \\ &= \frac{1}{6} \int_0^1 (s^3 + 3s^2 + 2s)\, ds \\ &= \frac{1}{6}\left[\frac{s^4}{4} + s^3 + s^2\right]_0^1 = \frac{1}{6}\left(\frac{9}{4}\right) = \frac{3}{8}.\end{aligned}$$

Por consiguiente,

$$\begin{aligned}\int_{t_i}^{t_{i+1}} f(t, y(t))\, dt &= h\left[f(t_i, y(t_i)) + \frac{1}{2}\nabla f(t_i, y(t_i)) + \frac{5}{12}\nabla^2 f(t_i, y(t_i)) + \cdots\right] \\ &\quad + \frac{h^{m+1}}{m!} \int_0^1 s(s+1)\cdots(s+m-1) f^{(m)}(\xi_i, y(\xi_i))\, ds. \end{aligned} \tag{5.29}$$

Puesto que $s(s+1)\cdots(s+m-1)$ no cambia de signo en [0, 1], el teorema de valor medio ponderado para integrales se puede usar para deducir que para algún número μ_i, donde $t_{i+1-m} < \mu_i < t_{i+1}$, el término de error en la ecuación (5.29) se convierte en

$$\frac{h^{m+1}}{m!}\int_0^1 s(s+1)\cdots(s+m-1)f^{(m)}(\xi_i, y(\xi_i))\,ds$$
$$= \frac{h^{m+1}f^{(m)}(\mu_i, y(\mu_i))}{m!}\int_0^1 s(s+1)\cdots(s+m-1)\,ds.$$

Por lo tanto, el error en la ecuación (5.29) se simplifica en

$$h^{m+1}f^{(m)}(\mu_i, y(\mu_i))(-1)^m\int_0^1 \binom{-s}{m}\,ds. \tag{5.30}$$

Sin embargo $y(t_{i+1}) - y(t_i) = \int_{t_i}^{t_{i+1}} f(t, y(t))\,dt$, por lo que la ecuación (5.27) se puede escribir como

$$y(t_{i+1}) = y(t_i) + h\left[f(t_i, y(t_i)) + \frac{1}{2}\nabla f(t_i, y(t_i)) + \frac{5}{12}\nabla^2 f(t_i, y(t_i)) + \cdots\right]$$
$$+ h^{m+1}f^{(m)}(\mu_i, y(\mu_i))(-1)^m\int_0^1 \binom{-s}{m}\,ds. \tag{5.31}$$

Ejemplo 2 Use la ecuación (5.31) con $m = 3$ para derivar la técnica de Adams-Bashforth de tres pasos.

Solución Tenemos

$$y(t_{i+1}) \approx y(t_i) + h\left[f(t_i, y(t_i)) + \frac{1}{2}\nabla f(t_i, y(t_i)) + \frac{5}{12}\nabla^2 f(t_i, y(t_i))\right]$$
$$= y(t_i) + h\left\{f(t_i, y(t_i)) + \frac{1}{2}[f(t_i, y(t_i)) - f(t_{i-1}, y(t_{i-1}))]\right.$$
$$\left.+ \frac{5}{12}[f(t_i, y(t_i)) - 2f(t_{i-1}, y(t_{i-1})) + f(t_{i-2}, y(t_{i-2}))]\right\}$$
$$= y(t_i) + \frac{h}{12}[23f(t_i, y(t_i)) - 16f(t_{i-1}, y(t_{i-1})) + 5f(t_{i-2}, y(t_{i-2}))].$$

El método Adams-Bashforth de tres pasos es, por consiguiente,

$$w_0 = \alpha, \quad w_1 = \alpha_1, \quad w_2 = \alpha_2,$$
$$w_{i+1} = w_i + \frac{h}{12}[23f(t_i, w_i) - 16f(t_{i-1}, w_{i-1})] + 5f(t_{i-2}, w_{i-2})],$$

para $i = 2, 3, \ldots, N-1$. ■

Los métodos multipasos también se pueden derivar por medio de la serie de Taylor. Un ejemplo del procedimiento implicado se considera en el ejercicio 17. Una derivación por medio del polinomio de interpolación de Lagrange se analiza en el ejercicio 16.

El error de truncamiento local para métodos multipasos se define de manera análoga al del método de un paso. En el caso de los métodos de un paso, el error de truncamiento local provee una medida de la forma en la que la solución de la ecuación diferencial no logra resolver la ecuación de diferencia.

Definición 5.15 Si $y(t)$ es la solución al problema de valor inicial

$$y' = f(t, y), \quad a \le t \le b, \quad y(a) = \alpha,$$

y

$$w_{i+1} = a_{m-1}w_i + a_{m-2}w_{i-1} + \cdots + a_0 w_{i+1-m}$$
$$+ h[b_m f(t_{i+1}, w_{i+1}) + b_{m-1} f(t_i, w_i) + \cdots + b_0 f(t_{i+1-m}, w_{i+1-m})]$$

es el $(i+1)$-ésimo paso en un método multipasos, el **error de truncamiento local** en este paso es

$$\tau_{i+1}(h) = \frac{y(t_{i+1}) - a_{m-1}y(t_i) - \cdots - a_0 y(t_{i+1-m})}{h} - [b_m f(t_{i+1}, y(t_{i+1})) + \cdots + b_0 f(t_{i+1-m}, y(t_{i+1-m}))], \tag{5.32}$$

para cada $i = m-1, m, \ldots, N-1$. ■

Ejemplo 3 Determine el error de truncamiento local para el método Adams-Bashforth de tres pasos derivado en el ejemplo 2.

Solución Al considerar la forma del error provisto en la ecuación (5.30), la entrada adecuada en la tabla 5.12, da

$$h^4 f^{(3)}(\mu_i, y(\mu_i))(-1)^3 \int_0^1 \binom{-s}{3} ds = \frac{3h^4}{8} f^{(3)}(\mu_i, y(\mu_i)).$$

A través del hecho de que $f^{(3)}(\mu_i, y(\mu_i)) = y^{(4)}(\mu_i)$ y la ecuación de diferencia derivada en el ejemplo 2, tenemos

$$\tau_{i+1}(h) = \frac{y(t_{i+1}) - y(t_i)}{h} - \frac{1}{12}[23f(t_i, y(t_i)) - 16f(t_{i-1}, y(t_{i-1})) + 5f(t_{i-2}, y(t_{i-2}))]$$
$$= \frac{1}{h}\left[\frac{3h^4}{8} f^{(3)}(\mu_i, y(\mu_i))\right] = \frac{3h^3}{8} y^{(4)}(\mu_i), \quad \text{para algunos } \mu_i \in (t_{i-2}, t_{i+1}).$$ ■

Métodos explícitos de Adams-Bashforth

Algunos de los métodos explícitos de Adams-Bashforth, junto con sus valores iniciales y errores de truncamiento local requeridos, son los siguientes. La derivación de estas técnicas es similar al procedimiento en los ejemplos 2 y 3.

Método explícito de dos pasos de Adams-Bashforth

$$w_0 = \alpha, \quad w_1 = \alpha_1,$$
$$w_{i+1} = w_i + \frac{h}{2}[3f(t_i, w_i) - f(t_{i-1}, w_{i-1})], \tag{5.33}$$

donde $i = 1, 2, \ldots, N-1$. El error de truncamiento local es $\tau_{i+1}(h) = \frac{5}{12} y'''(\mu_i)h^2$, para algunos $\mu_i \in (t_{i-1}, t_{i+1})$.

Método explícito de tres pasos de Adams-Bashforth

$$w_0 = \alpha, \quad w_1 = \alpha_1, \quad w_2 = \alpha_2,$$
$$w_{i+1} = w_i + \frac{h}{12}[23f(t_i, w_i) - 16f(t_{i-1}, w_{i-1}) + 5f(t_{i-2}, w_{i-2})], \tag{5.34}$$

donde $i = 2, 3, \ldots, N-1$. El error de truncamiento local es $\tau_{i+1}(h) = \frac{3}{8} y^{(4)}(\mu_i)h^3$, para algunos $\mu_i \in (t_{i-2}, t_{i+1})$.

Método explícito de cuatro pasos de Adams-Bashforth

$$w_0 = \alpha, \quad w_1 = \alpha_1, \quad w_2 = \alpha_2, \quad w_3 = \alpha_3,$$

$$w_{i+1} = w_i + \frac{h}{24}[55f(t_i, w_i) - 59f(t_{i-1}, w_{i-1}) + 37f(t_{i-2}, w_{i-2}) - 9f(t_{i-3}, w_{i-3})], \tag{5.35}$$

donde $i = 3, 4, \ldots, N-1$. El error de truncamiento local es $\tau_{i+1}(h) = \frac{251}{720}y^{(5)}(\mu_i)h^4$, para algunos $\mu_i \in (t_{i-3}, t_{i+1})$.

Método explícito de cinco pasos de Adams-Bashforth

$$w_0 = \alpha, \quad w_1 = \alpha_1, \quad w_2 = \alpha_2, \quad w_3 = \alpha_3, \quad w_4 = \alpha_4,$$

$$w_{i+1} = w_i + \frac{h}{720}[1901f(t_i, w_i) - 2774f(t_{i-1}, w_{i-1})$$
$$+ 2616f(t_{i-2}, w_{i-2}) - 1274f(t_{i-3}, w_{i-3}) + 251f(t_{i-4}, w_{i-4})], \tag{5.36}$$

donde $i = 4, 5, \ldots, N-1$. El error de truncamiento local es $\tau_{i+1}(h) = \frac{95}{288}y^{(6)}(\mu_i)h^5$, para algunos $\mu_i \in (t_{i-4}, t_{i+1})$.

Métodos implícitos de Adams-Moulton

Los métodos implícitos se derivan a través de $(t_{i+1}, f(t_{i+1}, y(t_{i+1})))$ como un nodo de interpolación adicional en la aproximación de la integral

$$\int_{t_i}^{t_{i+1}} f(t, y(t))\, dt.$$

Algunos de los métodos implícitos más comunes son los siguientes.

Métodos implícitos de dos pasos de Adams-Moulton

$$w_0 = \alpha, \quad w_1 = \alpha_1,$$

$$w_{i+1} = w_i + \frac{h}{12}[5f(t_{i+1}, w_{i+1}) + 8f(t_i, w_i) - f(t_{i-1}, w_{i-1})], \tag{5.37}$$

donde $i = 1, 2, \ldots, N-1$. El error de truncamiento local es $\tau_{i+1}(h) = -\frac{1}{24}y^{(4)}(\mu_i)h^3$, para algunos $\mu_i \in (t_{i-1}, t_{i+1})$.

Métodos implícitos de tres pasos de Adams-Moulton

$$w_0 = \alpha, \quad w_1 = \alpha_1, \quad w_2 = \alpha_2,$$

$$w_{i+1} = w_i + \frac{h}{24}[9f(t_{i+1}, w_{i+1}) + 19f(t_i, w_i) - 5f(t_{i-1}, w_{i-1}) + f(t_{i-2}, w_{i-2})], \tag{5.38}$$

donde $i = 2, 3, \ldots, N-1$. El error de truncamiento local es $\tau_{i+1}(h) = -\frac{19}{720}y^{(5)}(\mu_i)h^4$, para algunos $\mu_i \in (t_{i-2}, t_{i+1})$.

Métodos implícitos de cuatro pasos de Adams-Moulton

$$w_0 = \alpha, \quad w_1 = \alpha_1, \quad w_2 = \alpha_2, \quad w_3 = \alpha_3,$$

$$w_{i+1} = w_i + \frac{h}{720}[251 f(t_{i+1}, w_{i+1}) + 646 f(t_i, w_i) - 264 f(t_{i-1}, w_{i-1})$$

$$+ 106 f(t_{i-2}, w_{i-2}) - 19 f(t_{i-3}, w_{i-3})], \tag{5.39}$$

donde $i = 3, 4, \ldots, N-1$. El error de truncamiento local es $\tau_{i+1}(h) = -\frac{3}{160} y^{(6)}(\mu_i)h^5$, para algunos $\mu_i \in (t_{i-3}, t_{i+1})$.

Es interesante comparar un método explícito de Adams-Bashforth de *m* pasos con un método implícito de Adams-Moulton de $(m-1)$ pasos. Ambos implican evaluaciones de *f* por paso y ambos tienen términos $y^{(m+1)}(\mu_i)h^m$ en sus errores de truncamiento local. En general, los coeficientes de los términos relacionados con *f* en el error de truncamiento local son más pequeños para los métodos implícitos que para los explícitos. Esto conduce a una mayor estabilidad y errores de redondeo más pequeños para los métodos implícitos.

Ejemplo 4 Considere el problema de valor inicial

$$y' = y - t^2 + 1, \quad 0 \le t \le 2, \quad y(0) = 0.5.$$

Utilice los valores exactos proporcionados por $y(t) = (t+1)^2 - 0.5e^t$ como valores iniciales y $h = 0.2$ para comparar las aproximaciones a partir de **a)** el método explícito de cuatro pasos de Adams-Bashforth y **b)** el método implícito de tres pasos de Adams-Moulton.

Solución **a)** El método Adams-Bashforth tiene la ecuación de diferencia

$$w_{i+1} = w_i + \frac{h}{24}[55 f(t_i, w_i) - 59 f(t_{i-1}, w_{i-1}) + 37 f(t_{i-2}, w_{i-2}) - 9 f(t_{i-3}, w_{i-3})],$$

para $i = 3, 4, \ldots, 9$. Al simplificar mediante $f(t, y) = y - t^2 + 1, h = 0.2$, y $t_i = 0.2i$, se convierte en

$$w_{i+1} = \frac{1}{24}[35w_i - 11.8w_{i-1} + 7.4w_{i-2} - 1.8w_{i-3} - 0.192i^2 - 0.192i + 4.736].$$

b) El método Adams-Moulton tiene la ecuación de diferencia

$$w_{i+1} = w_i + \frac{h}{24}[9 f(t_{i+1}, w_{i+1}) + 19 f(t_i, w_i) - 5 f(t_{i-1}, w_{i-1}) + f(t_{i-2}, w_{i-2})],$$

para $i = 2, 3, \ldots, 9$. Esto se reduce a

$$w_{i+1} = \frac{1}{24}[1.8w_{i+1} + 27.8w_i - w_{i-1} + 0.2w_{i-2} - 0.192i^2 - 0.192i + 4.736].$$

Para usar este método de manera explícita, necesitamos resolver explícitamente la ecuación para w_{i+1}. Esto nos da

$$w_{i+1} = \frac{1}{22.2}[27.8w_i - w_{i-1} + 0.2w_{i-2} - 0.192i^2 - 0.192i + 4.736],$$

para $i = 2, 3, \ldots, 9$.

Los resultados en la tabla 5.13 se obtuvieron a partir de valores exactos de $y(t) = (t+1)^2 - 0.5e^t$ para α, α_1, α_2 y α_3 en el caso explícito de Adams-Bashforth y para α, α_1 y α_2 en el caso implícito de Adams-Moulton. Observe que el método Adams Moulton proporciona resultados consistentemente mejores. ■

Tabla 5.13

t_i	Exacto	Adams-Bashforth w_i	Error	Adams-Moulton w_i	Error
0.0	0.5000000				
0.2	0.8292986				
0.4	1.2140877				
0.6	1.6489406			1.6489341	0.0000065
0.8	2.1272295	2.1273124	0.0000828	2.1272136	0.0000160
1.0	2.6408591	2.6410810	0.0002219	2.6408298	0.0000293
1.2	3.1799415	3.1803480	0.0004065	3.1798937	0.0000478
1.4	3.7324000	3.7330601	0.0006601	3.7323270	0.0000731
1.6	4.2834838	4.2844931	0.0010093	4.2833767	0.0001071
1.8	4.8151763	4.8166575	0.0014812	4.8150236	0.0001527
2.0	5.3054720	5.3075838	0.0021119	5.3052587	0.0002132

Métodos indicador-corrector

En el ejemplo 4, el método Adams–Moulton da mejores resultados que el método explícito de Adams–Bashforth del mismo orden. A pesar de que, en general, es el caso, los métodos implícitos tienen la debilidad inherente de tener que convertir primero el método de manera algebraica para una representación explícita de w_{i+1}. Este procedimiento no siempre es posible, como se observa al considerar el problema fundamental de valor inicial

$$y' = e^y, \quad 0 \leq t \leq 0.25, \quad y(0) = 1.$$

Puesto que $f(t, y) = e^y$, el método de tres pasos de Adams–Moulton tiene

$$w_{i+1} = w_i + \frac{h}{24}[9e^{w_{i+1}} + 19e^{w_i} - 5e^{w_{i-1}} + e^{w_{i-2}}]$$

como ecuación de diferencia y esta ecuación no se puede resolver de manera algebraica para w_{i+1}.

Podríamos usar el método de Newton o el método de secante para aproximar w_{i+1}, pero esto complica considerablemente el procedimiento. En la práctica, los métodos implícitos multipasos no se utilizan de acuerdo con lo descrito antes. Más bien, se usan para mejorar las aproximaciones obtenidas con los métodos explícitos. La combinación de un método explícito para predecir y uno implícito para mejorar la predicción recibe el nombre de **método indicador-corrector**.

Considere el siguiente método de cuarto orden para resolver un problema de valor inicial. El primer paso es calcular los valores iniciales w_0, w_1, w_2 y w_3 para el método de cuatro pasos de Adams–Bashforth. Para hacerlo usamos el método de un paso de cuarto orden, el método Runge-Kutta de cuarto orden. Lo siguiente es calcular una aproximación w_{4p}, para $y(t_4)$ por medio del método explícito de Adams-Bashforth como indicador:

$$w_{4p} = w_3 + \frac{h}{24}[55f(t_3, w_3) - 59f(t_2, w_2) + 37f(t_1, w_1) - 9f(t_0, w_0)].$$

Esta aproximación mejora al insertar w_{4p} en el lado derecho del método implícito de tres pasos de Adams–Moulton y usar ese método como corrector. Esto nos da

$$w_4 = w_3 + \frac{h}{24}[9f(t_4, w_{4p}) + 19f(t_3, w_3) - 5f(t_2, w_2) + f(t_1, w_1)].$$

La única evaluación de función nueva requerida en este procedimiento es $f(t_4, w_{4p})$ en la ecuación de corrección; todos los demás valores de f se han calculado para aproximaciones previas.

A continuación se usa el valor w_4 como la aproximación para $y(t_4)$, y la técnica para utilizar el método Adams–Bashforth como indicador y el método Adams–Moulton como corrector se repite para encontrar w_{5p} y w_5, las aproximaciones inicial y final para $y(t_5)$. Este proceso continúa hasta que obtenemos una aproximación w_N para $y(t_N) = y(b)$.

Las aproximaciones mejoradas para $y(t_{i+1})$ se pueden obtener al iterar la fórmula de Adams–Moulton, pero esto converge para la aproximación provista por la fórmula implícita en lugar de la solución $y(t_{i+1})$. Por lo tanto, normalmente, es más eficiente usar una reducción en el tamaño de paso si se necesita precisión mejorada.

El algoritmo 5.4 está basado en el método Adams-Bashforth de cuarto orden como indicador y una iteración para el método Adams–Moulton como corrector, con los valores iniciales obtenidos a partir del método Runge-Kutta de cuarto orden.

ALGORITMO **5.4**

Indicador-corrector de Adams de cuarto orden

Para aproximar la solución del problema de valor inicial

$$y' = f(t, y), \quad a \le t \le b, \quad y(a) = \alpha,$$

en $(N+1)$ números igualmente espaciados en el intervalo $[a, b]$:

ENTRADA extremos a, b; entero N; condición inicial α.

SALIDA aproximación w para y en los valores $(N+1)$ de t.

Paso 1 Determine $h = (b-a)/N$;
$t_0 = a$;
$w_0 = \alpha$;
SALIDA (t_0, w_0).

Paso 2 Para $i = 1, 2, 3$, haga los pasos 3–5.
(*Calcule los valores iniciales con el método Runge-Kutta*.)

Paso 3 Determine $K_1 = hf(t_{i-1}, w_{i-1})$;
$K_2 = hf(t_{i-1} + h/2, w_{i-1} + K_1/2)$;
$K_3 = hf(t_{i-1} + h/2, w_{i-1} + K_2/2)$;
$K_4 = hf(t_{i-1} + h, w_{i-1} + K_3)$.

Paso 4 Determine $w_i = w_{i-1} + (K_1 + 2K_2 + 2K_3 + K_4)/6$;
$t_i = a + ih$.

Paso 5 SALIDA (t_i, w_i).

Paso 6 Para $i = 4, \ldots, N$ haga los pasos 7–10.

Paso 7 Determine $t = a + ih$;
$w = w_3 + h[55f(t_3, w_3) - 59f(t_2, w_2) + 37f(t_1, w_1) - 9f(t_0, w_0)]/24$; (*Prediga* w_i.)
$w = w_3 + h[9f(t, w) + 19f(t_3, w_3) - 5f(t_2, w_2) + f(t_1, w_1)]/24$. (*Corrija* w_i.)

Paso 8 SALIDA (t, w).

Paso 9 Para $j = 0, 1, 2$
Determine $t_j = t_{j+1}$; (*Prepara la siguiente iteración.*)
$w_j = w_{j+1}$.

Paso 10 Determine $t_3 = t$;
$w_3 = w$.

Paso 11 PARE. ■

Ejemplo 5 Aplique el método indicador-corrector de cuarto orden de Adams con $h = 0.2$ y los valores iniciales del método de cuarto orden de Runge-Kutta para el problema de valor inicial

$$y' = y - t^2 + 1, \quad 0 \le t \le 2, \quad y(0) = 0.5.$$

Solución Ésta es una continuación y modificación del problema considerado en el ejemplo 1 al inicio de la sección. En ese ejemplo, encontramos que las aproximaciones de inicio a partir de Runge-Kutta son

$$y(0) = w_0 = 0.5, \; y(0.2) \approx w_1 = 0.8292933, \; y(0.4) \approx w_2 = 1.2140762, \; \text{y} \; y(0.6) \approx w_3 = 1.6489220,$$

y el método Adams–Bashforth de cuarto orden nos da

$$\begin{aligned}
y(0.8) \approx w_{4p} &= w_3 + \frac{0.2}{24}(55f(0.6, w_3) - 59f(0.4, w_2) + 37f(0.2, w_1) - 9f(0, w_0)) \\
&= 1.6489220 + \frac{0.2}{24}(55f(0.6, 1.6489220) - 59f(0.4, 1.2140762) \\
&\quad + 37f(0.2, 0.8292933) - 9f(0, 0.5)) \\
&= 1.6489220 + 0.0083333(55(2.2889220) - 59(2.0540762) \\
&\quad + 37(1.7892933) - 9(1.5)) \\
&= 2.1272892.
\end{aligned}$$

Ahora utilizaremos w_{4p} como indicador de la aproximación para $y(0.8)$ y determinaremos el valor correcto w_4, a partir del método implícito de Adams–Moulton. Esto nos da

$$\begin{aligned}
y(0.8) \approx w_4 &= w_3 + \frac{0.2}{24}\left(9f(0.8, w_{4p}) + 19f(0.6, w_3) - 5f(0.4, w_2) + f(0.2, w_1)\right) \\
&= 1.6489220 + \frac{0.2}{24}(9f(0.8, 2.1272892) + 19f(0.6, 1.6489220) \\
&\quad - 5f(0.4, 1.2140762) + f(0.2, 0.8292933)) \\
&= 1.6489220 + 0.0083333(9(2.4872892) + 19(2.2889220) - 5(2.0540762) \\
&\quad + (1.7892933)) \\
&= 2.1272056.
\end{aligned}$$

Ahora utilizamos esta aproximación para determinar el indicador w_{5p}, para $y(1.0)$ como

$$\begin{aligned}
y(1.0) \approx w_{5p} &= w_4 + \frac{0.2}{24}(55f(0.8, w_4) - 59f(0.6, w_3) + 37f(0.4, w_2) - 9f(0.2, w_1)) \\
&= 2.1272056 + \frac{0.2}{24}(55f(0.8, 2.1272056) - 59f(0.6, 1.6489220) \\
&\quad + 37f(0.4, 1.2140762) - 9f(0.2, 0.8292933)) \\
&= 2.1272056 + 0.0083333(55(2.4872056) - 59(2.2889220) \\
&\quad + 37(2.0540762) - 9(1.7892933)) \\
&= 2.6409314
\end{aligned}$$

y corregimos esto con

$$y(1.0) \approx w_5 = w_4 + \frac{0.2}{24}\left(9f(1.0, w_{5p}) + 19f(0.8, w_4) - 5f(0.6, w_3) + f(0.4, w_2)\right)$$

$$= 2.1272056 + \frac{0.2}{24}(9f(1.0, 2.6409314) + 19f(0.8, 2.1272892)$$

$$- 5f(0.6, 1.6489220) + f(0.4, 1.2140762))$$

$$= 2.1272056 + 0.0083333(9(2.6409314) + 19(2.4872056) - 5(2.2889220)$$

$$+ (2.0540762))$$

$$= 2.6408286.$$

En el ejemplo 1, encontramos que utilizar el método explícito de Adams–Bashforth sólo produjo resultados que eran inferiores a los de Runge-Kutta. Sin embargo, estas aproximaciones para $y(0.8)$ y $y(1.0)$ son precisas, respectivamente, dentro de

$$|2.1272295 - 2.1272056| = 2.39 \times 10^{-5} \text{ y } |2.6408286 - 2.6408591| = 3.05 \times 10^{-5},$$

en comparación con las de Runge-Kutta, lo cual es preciso, respectivamente, dentro de

$$|2.1272027 - 2.1272892| = 2.69 \times 10^{-5} \text{ y } |2.6408227 - 2.6408591| = 3.64 \times 10^{-5}.$$

Las aproximaciones restantes del indicador-corrector se generaron mediante el algoritmo 5.4 y se muestran en la tabla 5.14. ■

Tabla 5.14

t_i	$y_i = y(t_i)$	w_i	Error $\lvert y_i - w_i \rvert$
0.0	0.5000000	0.5000000	0
0.2	0.8292986	0.8292933	0.0000053
0.4	1.2140877	1.2140762	0.0000114
0.6	1.6489406	1.6489220	0.0000186
0.8	2.1272295	2.1272056	0.0000239
1.0	2.6408591	2.6408286	0.0000305
1.2	3.1799415	3.1799026	0.0000389
1.4	3.7324000	3.7323505	0.0000495
1.6	4.2834838	4.2834208	0.0000630
1.8	4.8151763	4.8150964	0.0000799
2.0	5.3054720	5.3053707	0.0001013

Edward Arthur Milne (1896–1950) trabajó en investigación balística durante la Primera Guerra Mundial y, después, para Solar Physics Observatory, en Cambridge. En 1929, fue nombrado W. W. Rouse Ball Chair en el Wadham College en Oxford.

Otros métodos multipasos se pueden derivar por medio de integración de polinomios de interpolación sobre intervalos de la forma $[t_j, t_{i+1}]$, para $j \leq i - 1$, para obtener una aproximación para $y(t_{i+1})$. Cuando se integra un polinomio de interpolación sobre $[t_{i-3}, t_{i+1}]$, el resultado es el **método explícito de Milne**:

$$w_{i+1} = w_{i-3} + \frac{4h}{3}[2f(t_i, w_i) - f(t_{i-1}, w_{i-1}) + 2f(t_{i-2}, w_{i-2})],$$

que tiene un error de truncamiento local $\frac{14}{45}h^4 y^{(5)}(\xi_i)$, para algunos $\xi_i \in (t_{i-3}, t_{i+1})$.

El nombre de Simpson está asociado con esta técnica porque se basa en la regla para la integración de Simpson.

El método de Milne se usa ocasionalmente como indicador para el **método implícito de Simpson**,

$$w_{i+1} = w_{i-1} + \frac{h}{3}[f(t_{i+1}, w_{i+1}) + 4f(t_i, w_i) + f(t_{i-1}, w_{i-1})],$$

que tiene un error de truncamiento local $-(h^4/90)y^{(5)}(\xi_i)$, para algunos $\xi_i \in (t_{i-1}, t_{i+1})$, y se obtiene al integrar un polinomio de interpolación sobre $[t_{i-1}, t_{i+1}]$.

En general, el error de truncamiento local relacionado con el método indicador corrector del tipo Milne-Simpson es más pequeño que el del método Adams–Bashforth-Moulton. Pero la técnica tiene uso limitado debido a los problemas de error de redondeo, los cuales no se presentan con el procedimiento de Adams. Esta elaboración se estudia en la sección 5.10.

La sección Conjunto de ejercicios 5.6 está disponible en línea. Encuentre la ruta de acceso en las páginas preliminares.

5.7 Método multipasos de tamaño de paso variable

El método Runge-Kutta-Fehlberg se utiliza para el control de error porque en cada paso ofrece, con un costo adicional pequeño, *dos* aproximaciones que se pueden comparar y relacionar con el error de truncamiento local. Las técnicas de indicador-corrector generan dos aproximaciones en cada paso, por lo que son candidatos naturales para la adaptación de control de error.

Para demostrar el procedimiento de control de error, construimos un método indicador-corrector de tamaño de paso variable mediante el método explícito Adams-Bashforth de cuatro pasos como indicador y el método implícito Adams-Moulton de tres pasos como corrector.

El método de Adams-Bashforth de cuatro pasos proviene de la relación

$$y(t_{i+1}) = y(t_i) + \frac{h}{24}[55f(t_i, y(t_i)) - 59f(t_{i-1}, y(t_{i-1})) + 37f(t_{i-2}, y(t_{i-2})) - 9f(t_{i-3}, y(t_{i-3}))] + \frac{251}{720}y^{(5)}(\hat{\mu}_i)h^5,$$

para algunos $\hat{\mu}_i \in (t_{i-3}, t_{i+1})$. La suposición de que las aproximaciones $w_0, w_1, \ldots, w_i$ son todas exactas, implica que el error de truncamiento local Adams-Bashforth es

$$\frac{y(t_{i+1}) - w_{p,i+1}}{h} = \frac{251}{720}y^{(5)}(\hat{\mu}_i)h^4. \tag{5.40}$$

Un análisis similar del método Adams-Moulton de tres pasos, que proviene de

$$y(t_{i+1}) = y(t_i) + \frac{h}{24}[9f(t_{i+1}, y(t_{i+1})) + 19f(t_i, y(t_i)) - 5f(t_{i-1}, y(t_{i-1})) + f(t_{i-2}, y(t_{i-2}))] - \frac{19}{720}y^{(5)}(\tilde{\mu}_i)h^5,$$

para algunos $\tilde{\mu}_i \in (t_{i-2}, t_{i+1})$, conduce al error de truncamiento local

$$\frac{y(t_{i+1}) - w_{i+1}}{h} = -\frac{19}{720}y^{(5)}(\tilde{\mu}_i)h^4. \tag{5.41}$$

Para avanzar más, debemos suponer que para valores pequeños de h, tenemos

$$y^{(5)}(\hat{\mu}_i) \approx y^{(5)}(\tilde{\mu}_i).$$

La efectividad de la técnica de control de error depende directamente de esta suposición.

Si restamos la ecuación (5.41) de la ecuación (5.40), tenemos

$$\frac{w_{i+1} - w_{p,i+1}}{h} = \frac{h^4}{720}[251y^{(5)}(\hat{\mu}_i) + 19y^{(5)}(\tilde{\mu}_i)] \approx \frac{3}{8}h^4 y^{(5)}(\tilde{\mu}_i),$$

por lo que

$$y^{(5)}(\tilde{\mu}_i) \approx \frac{8}{3h^5}(w_{i+1} - w_{p,i+1}). \tag{5.42}$$

El uso de este resultado para eliminar el término relacionado con $y^{(5)}(\tilde{\mu}_i)h^4$ a partir de la ecuación (5.41) provee la aproximación para el error de truncamiento local Adams-Moulton

$$|\tau_{i+1}(h)| = \frac{|y(t_{i+1}) - w_{i+1}|}{h} \approx \frac{19h^4}{720} \cdot \frac{8}{3h^5}|w_{i+1} - w_{p,i+1}| = \frac{19|w_{i+1} - w_{p,i+1}|}{270h}.$$

Suponga que ahora reconsideramos (ecuación 5.41) con un nuevo tamaño de paso qh al generar aproximaciones nuevas $\hat{w}_{p,i+1}$ y $\hat{w}_{i+1}$. El objetivo es seleccionar q de tal forma que el error de truncamiento local provisto en la ecuación (5.41) esté acotada por una tolerancia prescrita ε. Si suponemos que el valor $y^{(5)}(\mu)$ en la ecuación (5.41) relacionado con qh también se aproxima mediante la ecuación (5.42), entonces

$$\begin{aligned}\frac{|y(t_i + qh) - \hat{w}_{i+1}|}{qh} &= \frac{19q^4h^4}{720}|y^{(5)}(\mu)| \approx \frac{19q^4h^4}{720}\left[\frac{8}{3h^5}|w_{i+1} - w_{p,i+1}|\right] \\ &= \frac{19q^4}{270}\frac{|w_{i+1} - w_{p,i+1}|}{h},\end{aligned}$$

y necesitamos seleccionar q de tal forma que

$$\frac{|y(t_i + qh) - \hat{w}_{i+1}|}{qh} \approx \frac{19q^4}{270}\frac{|w_{i+1} - w_{p,i+1}|}{h} < \varepsilon.$$

Es decir, seleccione q de tal forma que

$$q < \left(\frac{270}{19}\frac{h\varepsilon}{|w_{i+1} - w_{p,i+1}|}\right)^{1/4} \approx 2\left(\frac{h\varepsilon}{|w_{i+1} - w_{p,i+1}|}\right)^{1/4}.$$

En este desarrollo se ha realizado una serie de suposiciones de aproximación, por lo que en la práctica q se selecciona de manera conservadora, a menudo como

$$q = 1.5\left(\frac{h\varepsilon}{|w_{i+1} - w_{p,i+1}|}\right)^{1/4}.$$

Un cambio en el tamaño de paso para un método multipasos es más costoso en términos de evaluaciones de función que para un método de un paso porque se deben calcular valores iniciales igualmente espaciados. En consecuencia, es una práctica común ignorar el cambio de tamaño de paso siempre que el error de truncamiento local esté entre $\varepsilon/10$ y ε, es decir, cuando

$$\frac{\varepsilon}{10} < |\tau_{i+1}(h)| = \frac{|y(t_{i+1}) - w_{i+1}|}{h} \approx \frac{19|w_{i+1} - w_{p,i+1}|}{270h} < \varepsilon.$$

Además, a q suele asignársele una cota superior para garantizar que una aproximación precisa única poco común no resulte en un tamaño de paso demasiado grande. El algoritmo 5.5 incluye esta protección con una cota superior de 4.

Recuerde que los métodos multipasos requieren tamaños de paso iguales para los valores iniciales. Por lo que cualquier cambio en el tamaño de paso necesita recalcular de nuevo los valores iniciales en ese punto. En los pasos 3, 16 y 19 del algoritmo 5.5, esto se hace llamando un subalgoritmo Runge-Kutta (algoritmo 5.2), configurado en el paso 1.

ALGORITMO
5.5

Indicador-corrector de tamaño de paso variable Adams

Para aproximar la solución del problema de valor inicial

$$y' = f(t, y), \quad a \le t \le b, \quad y(a) = \alpha$$

con error de truncamiento local dentro de una tolerancia determinada:

ENTRADA extremos a, b; condición inicial α; tolerancia TOL; tamaño máximo de paso $hmáx$; tamaño mínimo de paso $hmín$.

SALIDA i, t_i, w_i, h, donde en el i-ésimo paso w_i se aproxima a $y(t_i)$ y se utiliza el tamaño de paso h, o un mensaje de que se excedió el tamaño mínimo de paso.

Paso 1 Configurar un subalgoritmo para el método Runge-Kutta de cuarto orden llamado a $RK4(h, v_0, x_0, v_1, x_1, v_2, x_2, v_3, x_3)$ que acepta como entrada un tamaño de paso h y valores iniciales $v_0 \approx y(x_0)$ y regresa $\{(x_j, v_j) \mid j = 1, 2, 3\}$ definidos mediante lo siguiente:

para $j = 1, 2, 3$

determine $K_1 = hf(x_{j-1}, v_{j-1})$;

$K_2 = hf(x_{j-1} + h/2, v_{j-1} + K_1/2)$

$K_3 = hf(x_{j-1} + h/2, v_{j-1} + K_2/2)$

$K_4 = hf(x_{j-1} + h, v_{j-1} + K_3)$

$v_j = v_{j-1} + (K_1 + 2K_2 + 2K_3 + K_4)/6$;

$x_j = x_0 + jh$.

Paso 2 Determine $t_0 = a$;

$w_0 = \alpha$;

$h = hmáx$;

$FLAG = 1$; (*FLAG se usará para salir del ciclo en el paso 4.*)

$LAST = 0$; (*LAST indicará cuando se calcula la última variable.*)

SALIDA (t_0, w_0).

Paso 3 Llame $RK4(h, w_0, t_0, w_1, t_1, w_2, t_2, w_3, t_3)$;

Determine $NFLAG = 1$; (*Indique el cálculo de RK 4.*)

$i = 4$;

$t = t_3 + h$.

Paso 4 Mientras $(FLAG = 1)$ haga los pasos 5–20.

Paso 5 Determine $WP = w_{i-1} + \frac{h}{24}[55f(t_{i-1}, w_{i-1}) - 59f(t_{i-2}, w_{i-2})$
$+ 37f(t_{i-3}, w_{i-3}) - 9f(t_{i-4}, w_{i-4})]$; (*Prediga w_i.*)

$WC = w_{i-1} + \frac{h}{24}[9f(t, WP) + 19f(t_{i-1}, w_{i-1})$
$- 5f(t_{i-2}, w_{i-2}) + f(t_{i-3}, w_{i-3})]$; (*Corrija w_i.*)

$\sigma = 19|WC - WP|/(270h)$.

Paso 6 Si $\sigma \le TOL$ entonces haga los pasos 7–16 (*Resultado aceptado.*)
también haga los pasos 17–19. (*Resultado rechazado.*)

Paso 7 Determine $w_i = WC$; (*Resultado aceptado.*)
$t_i = t$.

Paso 8 Si $NFLAG = 1$ entonces para $j = i-3, i-2, i-1, i$
SALIDA (j, t_j, w_j, h);
(*Resultados previos también aceptados.*)
también SALIDA (i, t_i, w_i, h).
(*Resultados previos ya aceptados.*)

Paso 9 Si $LAST = 1$ entonces determine $FLAG = 0$ (*El siguiente paso es* 20.)
también haga los pasos 10–16.

Paso 10 Determine $i = i + 1$;
$NFLAG = 0$.

Paso 11 Si $\sigma \leq 0.1\ TOL$ o $t_{i-1} + h > b$ entonces haga los pasos 12–16.
(*Incrementa h si es más preciso que lo requerido o disminuye h para incluir b como punto de malla*.)

Paso 12 Determine $q = (TOL/(2\sigma))^{1/4}$.

Paso 13 Si $q > 4$ entonces determine $h = 4h$
también determine $h = qh$.

Paso 14 Si $h > hmáx$ entonces determine $h = hmáx$.

Paso 15 Si $t_{i-1} + 4h > b$ entonces
determine $h = (b - t_{i-1})/4$;
$LAST = 1$.

Paso 16 Llame $RK4(h, w_{i-1}, t_{i-1}, w_i, t_i, w_{i+1}, t_{i+1}, w_{i+2}, t_{i+2})$;
Determine $NFLAG = 1$;
$i = i + 3$. (*Rama verdadera completada. Termina paso 6*.)
Próximo paso 20

Paso 17 Determine $q = (TOL/(2\sigma))^{1/4}$. (*Rama falsa a partir del paso 6: Resultado rechazado*.)

Paso 18 Si $q < 0.1$ entonces determine $h = 0.1h$
también determine $h = qh$.

Paso 19 Si $h < hmín$ entonces determine $FLAG = 0$;
SALIDA (‘*mínima excedida*’)
también
si $NFLAG = 1$ entonces determine $i = i - 3$;
(*Resultados previos también rechazados*.)
Llame $RK4(h, w_{i-1}, t_{i-1}, w_i, t_i, w_{i+1}, t_{i+1}, w_{i+2}, t_{i+2})$;
determine $i = i + 3$;
$NFLAG = 1$. (*Fin del paso 6*.)

Paso 20 Determine $t = t_{i-1} + h$. (*Fin del paso 4*.)

Paso 21 PARE. ■

Ejemplo 1 Utilice el método indicador-corrector de tamaño de paso variable de Adams con tamaño máximo de paso $hmáx = 0.2$, tamaño mínimo de paso $hmín = 0.01$, y tolerancia $TOL = 10^{-5}$ para aproximar la solución del problema de valor inicial

$$y' = y - t^2 + 1, \quad 0 \leq t \leq 2, \quad y(0) = 0.5.$$

Solución Comenzamos con $h = hmáx = 0.2$ y obtenemos w_0, w_1, w_2, y w_3 mediante Runge-Kutta, a continuación encontramos wp_4 y wc_4 al aplicar el método indicador-corrector. Estos cálculos se realizaron en el ejemplo 5 de la sección 5.6, donde se determinó que las aproximaciones Runge-Kutta con

$$y(0) = w_0 = 0.5,\ \ y(0.2) \approx w_1 = 0.8292933,\ \ y(0.4) \approx w_2 = 1.2140762, \ \ \text{y}$$
$$y(0.6) \approx w_3 = 1.6489220.$$

El indicador y corrector dan

$$y(0) = w_0 = 0.5, \;\; y(0.2) \approx w_1 = 0.8292933, \;\; y(0.4) \approx w_2 = 1.2140762, \quad \text{y}$$

$$y(0.6) \approx w_3 = 1.6489220$$

y

$$y(0.8) \approx w_{4p} = w_3 + \frac{0.2}{24}\,(55f(0.6, w_3) - 59f(0.4, w_2) + 37f(0.2, w_1) - 9f(0, w_0))$$
$$= 2.1272892$$

y

$$y(0.8) \approx w_{4c} = w_3 + \frac{0.2}{24}\big(9f(0.8, w_{4p}) + 19f(0.6, w_3) - 5f(0.42, w_2) + f(0.2, w_1)\big)$$
$$= 2.1272056.$$

Ahora necesitamos determinar si las aproximaciones son suficientemente precisas o si se necesita cambiar el tamaño de paso. Primero, encontramos

$$\sigma = \frac{19}{270h}|w_{4c} - w_{4p}| = \frac{19}{270(0.2)}|2.1272056 - 2.1272892| = 2.941 \times 10^{-5}.$$

Puesto que esto excede la tolerancia de 10^{-5}, se necesita un tamaño de paso nuevo y éste es

$$qh = \left(\frac{10^{-5}}{2\delta}\right)^{1/4} = \left(\frac{10^{-5}}{2(2.941 \times 10^{-5})}\right)^{1/4}(0.2) = 0.642(0.2) \approx 0.128.$$

Como consecuencia, necesitamos comenzar de nuevo con el procedimiento, al calcular los valores de Runge-Kutta con este tamaño de paso y, a continuación, utilizamos el método indicador-corrector con este tamaño de paso para calcular los valores nuevos de w_{4p} y w_{4c}. Después, necesitamos ejecutar una verificación de precisión en estas aproximaciones para ver si tuvimos éxito. La tabla 5.15 muestra si esta segunda ejecución es exitosa y enumera todos los resultados obtenidos en el algoritmo 5.5. ■

Tabla 5.15

t_i	$y(t_i)$	w_i	h_i	σ_i	$\lvert y(t_i) - w_i \rvert$
0	0.5	0.5			
0.12841297	0.70480460	0.70480402	0.12841297	4.431680×10^{-6}	0.0000005788
0.25682594	0.93320140	0.93320019	0.12841297	4.431680×10^{-6}	0.0000012158
0.38523891	1.18390410	1.18390218	0.12841297	4.431680×10^{-6}	0.0000019190
0.51365188	1.45545014	1.45544767	0.12841297	4.431680×10^{-6}	0.0000024670
0.64206485	1.74617653	1.74617341	0.12841297	5.057497×10^{-6}	0.0000031210
0.77047782	2.05419248	2.05418856	0.12841297	5.730989×10^{-6}	0.0000039170
0.89889079	2.37734803	2.37734317	0.12841297	6.522850×10^{-6}	0.0000048660
1.02730376	2.71319871	2.71319271	0.12841297	7.416639×10^{-6}	0.0000060010
1.15571673	3.05896505	3.05895769	0.12841297	8.433180×10^{-6}	0.0000073570
1.28412970	3.41148675	3.41147778	0.12841297	9.588365×10^{-6}	0.0000089720
1.38980552	3.70413577	3.70412572	0.10567582	7.085927×10^{-6}	0.0000100440
1.49548134	3.99668536	3.99667414	0.10567582	7.085927×10^{-6}	0.0000112120
1.60115716	4.28663498	4.28662249	0.10567582	7.085927×10^{-6}	0.0000124870
1.70683298	4.57120536	4.57119105	0.10567582	7.085927×10^{-6}	0.0000143120
1.81250880	4.84730747	4.84729107	0.10567582	7.844396×10^{-6}	0.0000163960
1.91818462	5.11150794	5.11148918	0.10567582	8.747367×10^{-6}	0.0000187650
1.93863847	5.16095461	5.16093546	0.02045384	1.376200×10^{-8}	0.0000191530
1.95909231	5.20978430	5.20976475	0.02045384	1.376200×10^{-8}	0.0000195490
1.97954616	5.25796697	5.25794701	0.02045384	1.376200×10^{-8}	0.0000199540
2.00000000	5.30547195	5.30545159	0.02045384	1.376200×10^{-8}	0.0000203670

La sección Conjunto de ejercicios 5.7 está disponible en línea. Encuentre la ruta de acceso en las páginas preliminares.



5.8 Métodos de extrapolación

La extrapolación se usó en la sección 4.5 para aproximar integrales definidas, en donde encontramos que al promediar de manera correcta las aproximaciones trapezoidales, relativamente imprecisas, se producían otras nuevas en extremo precisas. En esta sección aplicaremos la extrapolación para incrementar la precisión de las aproximaciones para la solución de problemas de valor inicial. Como observamos antes, las aproximaciones originales deben tener una expansión de error de una forma específica para que el procedimiento sea exitoso.

Para aplicar la extrapolación en la resolución de problemas de valor inicial usamos la técnica con base en el método de punto medio:

$$w_{i+1} = w_{i-1} + 2hf(t_i, w_i), \quad \text{para } i \geq 1. \tag{5.43}$$

Esta técnica requiere dos valores iniciales w_0 y w_1 que se necesitan antes de que se pueda determinar la primera aproximación del punto medio w_2. Un valor inicial es la condición inicial para $w_0 = y(a) = \alpha$. Para determinar el segundo valor inicial, w_1, aplicamos el método de Euler. Las aproximaciones subsiguientes se obtienen a partir de la ecuación (5.43). Después de generar una serie de aproximaciones de este tipo que terminan en un valor t, se realiza una corrección del extremo relacionada con las dos aproximaciones finales de punto medio. Esto produce una aproximación $w(t, h)$ para $y(t)$ que tiene la forma

$$y(t) = w(t, h) + \sum_{k=1}^{\infty} \delta_k h^{2k}, \tag{5.44}$$

donde δ_k son constantes relacionadas con las derivadas de la solución $y(t)$. El punto importante es que δ_k no depende del tamaño de paso h. Los detalles de este procedimiento se pueden encontrar en el artículo de Gragg [Gr].

Para ilustrar la técnica de extrapolación para resolver

$$y'(t) = f(t, y), \quad a \leq t \leq b, \quad y(a) = \alpha,$$

suponga que tenemos un tamaño de punto fijo h. Nos gustaría aproximar $y(t_1) = y(a + h)$.

Para el primer paso de extrapolación, dejamos que $h_0 = h/2$ y usamos el método de Euler con $w_0 = \alpha$ para aproximar $y(a + h_0) = y(a + h/2)$ como

$$w_1 = w_0 + h_0 f(a, w_0).$$

Aplicamos entonces el método de punto medio con $t_{i-1} = a$ y $t_i = a + h_0 = a + h/2$ para producir una primera aproximación $y(a + h) = y(a + 2h_0)$,

$$w_2 = w_0 + 2h_0 f(a + h_0, w_1).$$

Se aplica la corrección del extremo para obtener la aproximación final para $y(a+h)$ para el tamaño de paso h_0. Esto resulta en la aproximación $O(h_0^2)$ para $y(t_1)$.

$$y_{1,1} = \frac{1}{2}[w_2 + w_1 + h_0 f(a + 2h_0, w_2)].$$

Guardamos la aproximación $y_{1,1}$ y descartamos los resultados intermedios w_1 y w_2.

Para obtener la aproximación $y_{2,1}$, para $y(t_1)$, hacemos $h_1 = h/4$ y usamos el método de Euler con $w_0 = \alpha$ para obtener una aproximación para $y(a + h_1) = y(a + h/4)$, que llamaremos w_1:

$$w_1 = w_0 + h_1 f(a, w_0).$$

A continuación aproximamos $y(a+2h_1) = y(a+h/2)$ con w_2, $y(a+3h_1) = y(a+3h/4)$ con w_3, y w_4 para $y(a+4h_1) = y(t_1)$ con el método de punto medio:

$$w_2 = w_0 + 2h_1 f(a+h_1, w_1),$$
$$w_3 = w_1 + 2h_1 f(a+2h_1, w_2),$$

y

$$w_4 = w_2 + 2h_1 f(a+3h_1, w_3).$$

Ahora, se aplica la corrección del extremo para w_3 y w_4 para producir la aproximación mejorada $O(h_1^2)$ para $y(t_1)$,

$$y_{2,1} = \frac{1}{2}[w_4 + w_3 + h_1 f(a+4h_1, w_4)].$$

Debido a la forma el error provisto en la ecuación (5.44), las dos aproximaciones para $y(a+h)$ tienen la propiedad de que

$$y(a+h) = y_{1,1} + \delta_1 \left(\frac{h}{2}\right)^2 + \delta_2 \left(\frac{h}{2}\right)^4 + \cdots = y_{1,1} + \delta_1 \frac{h^2}{4} + \delta_2 \frac{h^4}{16} + \cdots$$

y

$$y(a+h) = y_{2,1} + \delta_1 \left(\frac{h}{4}\right)^2 + \delta_2 \left(\frac{h}{4}\right)^4 + \cdots = y_{2,1} + \delta_1 \frac{h^2}{16} + \delta_2 \frac{h^4}{256} + \cdots .$$

Podemos eliminar la parte $O(h^2)$ de este error de truncamiento al promediar las dos fórmulas de manera adecuada. Especialmente, si restamos la primera fórmula a cuatro veces la segunda y dividimos el resultado entre tres, obtenemos

$$y(a+h) = y_{2,1} + \frac{1}{3}(y_{2,1} - y_{1,1}) - \delta_2 \frac{h^4}{64} + \cdots .$$

Por lo que la aproximación para $y(t_1)$ dada por

$$y_{2,2} = y_{2,1} + \frac{1}{3}(y_{2,1} - y_{1,1})$$

tiene un error de orden $O(h^4)$.

A continuación hacemos $h_2 = h/6$ y aplicamos el método de Euler una vez, seguido del método de punto medio cinco veces. A continuación, utilizamos la corrección del extremo para determinar la aproximación h^2, $y_{3,1}$, para $y(a+h) = y(t_1)$. Esta aproximación se puede promediar con $y_{2,1}$ para producir una segunda aproximación $O(h^4)$ que designamos $y_{3,2}$. A continuación, se promedia $y_{3,2}$ y $y_{2,2}$ para eliminar los términos de error $O(h^4)$ y producimos una aproximación con error de orden $O(h^6)$. Se generan fórmulas de orden superior al continuar con el proceso.

La única diferencia significativa entre la extrapolación realizada aquí y la que utiliza integración de Romberg en la sección 4.5 resulta de la forma de seleccionar las subdivisiones. En la integración de Romberg, existe una fórmula conveniente para representar las aproximaciones de la regla trapezoidal compuesta que usa divisiones consecutivas del tamaño de paso mediante los enteros 1, 2, 4, 8, 16, 32, 64, ... Este procedimiento permite el proceso de promediado para continuar de una manera que se puede seguir fácilmente.

No tenemos los medios para producir fácilmente aproximaciones refinadas para problemas de valor inicial, por lo que se seleccionan las divisiones para la técnica de extrapolación para minimizar el número de evaluaciones de función requerido. El procedimiento de

promediado que surge de esta selección de la subdivisión, mostrado en la tabla 5.16, no es básico, pero, aparte de eso, el proceso es igual al que se usa para la integración de Romberg.

Tabla 5.16

$y_{1,1} = w(t, h_0)$		
$y_{2,1} = w(t, h_1)$	$y_{2,2} = y_{2,1} + \dfrac{h_1^2}{h_0^2 - h_1^2}(y_{2,1} - y_{1,1})$	
$y_{3,1} = w(t, h_2)$	$y_{3,2} = y_{3,1} + \dfrac{h_2^2}{h_1^2 - h_2^2}(y_{3,1} - y_{2,1})$	$y_{3,3} = y_{3,2} + \dfrac{h_2^2}{h_0^2 - h_2^2}(y_{3,2} - y_{2,2})$

El algoritmo 5.6 utiliza nodos de la forma 2^n y $2^n \cdot 3$. Es posible usar otras opciones.

El algoritmo 5.6 utiliza la técnica de extrapolación con la secuencia de enteros

$$q_0 = 2,\ q_1 = 4,\ q_2 = 6,\ q_3 = 8,\ q_4 = 12,\ q_5 = 16,\ q_6 = 24 \quad \text{y} \quad q_7 = 32.$$

Se selecciona un tamaño de paso básico h, y el método progresa mediante $h_i = h/q_i$, para cada $i = 0, \ldots, 7$, para aproximar $y(t+h)$. El error se controla al requerir que se calculen las aproximaciones $y_{1,1}, y_{2,2}, \ldots$ hasta $|y_{i,i} - y_{i-1,i-1}|$ sea menor a la tolerancia provista. Si la tolerancia no se alcanza mediante $i = 8$, entonces se disminuye h y el proceso se vuelve a aplicar.

Se especifican los valores máximo y mínimo de h, *hmín*, y *hmáx*, respectivamente, para garantizar el control del método. Si se encuentra que $y_{i,i}$ es aceptable, entonces w_1 se configura en $y_{i,i}$ y los cálculos comienzan de nuevo para determinar w_2, que aproximará $y(t_2) = y(a+2h)$. El proceso se repite hasta que se determina la aproximación w_N para $y(b)$.

ALGORITMO **5.6**

Extrapolación

Para aproximar la solución del problema de valor inicial

$$y' = f(t, y), \quad a \le t \le b, \quad y(a) = \alpha,$$

con error de truncamiento local dentro de una tolerancia determinada:

ENTRADA extremos a, b; condición inicial α; tolerancia TOL; tamaño máximo de paso *hmáx*; tamaño mínimo de paso *hmín*.

SALIDA T, W, h, donde W aproxima $y(t)$ y se usa el tamaño de paso h o un mensaje de que se excedió el tamaño mínimo de paso.

Paso 1 Inicialice el arreglo $NK = (2, 4, 6, 8, 12, 16, 24, 32)$.

Paso 2 Determine $TO = a$;
$WO = \alpha$;
$h = hmáx$;
$FLAG = 1$. (*Se usa FLAG para salir del ciclo en el paso* 4.)

Paso 3 Para $i = 1, 2, \ldots, 7$
para $j = 1, \ldots, i$
determine $Q_{i,j} = (NK_{i+1}/NK_j)^2$. (*Nota:* $Q_{i,j} = h_j^2/h_{i+1}^2$.)

Paso 4 Mientras ($FLAG = 1$) haga los pasos 5–20.

Paso 5 Determine $k = 1$;
$NFLAG = 0$. (*Cuando se alcanza la precisión deseada, NFLAG se configura en* 1.)

Paso 6 Mientras ($k \leq 8$ y $NFLAG = 0$) haga los pasos 7–14.

Paso 7 Determine $HK = h/NK_k$;
$T = TO$;
$W2 = WO$;
$W3 = W2 + HK \cdot f(T, W2)$; (*Primer paso de Euler.*)
$T = TO + HK$.

Paso 8 Para $j = 1, \ldots, NK_k - 1$
determine $W1 = W2$;
$W2 = W3$;
$W3 = W1 + 2HK \cdot f(T, W2)$; (*Método de punto medio.*)
$T = TO + (j + 1) \cdot HK$.

Paso 9 Determine $y_k = [W3 + W2 + HK \cdot f(T, W3)]/2$.
(*Corrección de extremo para calcular* $y_{k,1}$.)

Paso 10 Si $k \geq 2$ haga los pasos 11–13.
(*Nota:* $y_{k-1} \equiv y_{k-1,1}, y_{k-2} \equiv y_{k-2,2}, \ldots, y_1 \equiv y_{k-1,k-1}$ *ya que sólo se guarda la fila previa de la tabla.*)

Paso 11 Determine $j = k$;
$v = y_1$. (*Guarda* $y_{k-1,k-1}$.)

Paso 12 Mientras ($j \geq 2$) haga

determine $$y_{j-1} = y_j + \frac{y_j - y_{j-1}}{Q_{k-1,j-1} - 1};$$

(*Extrapolación para calcular* $y_{j-1} \equiv y_{k,k-j+2}$.)

$$\left(\textit{Nota:} \quad y_{j-1} = \frac{h_{j-1}^2 y_j - h_k^2 y_{j-1}}{h_{j-1}^2 - h_k^2}. \right)$$

$j = j - 1$.

Paso 13 Si $|y_1 - v| \leq TOL$ entonces haga $NFLAG = 1$.
(y_1 *aceptado como w nueva.*)

Paso 14 Determine $k = k + 1$. (*Termina paso 6*)

Paso 15 Determine $k = k - 1$. (*Parte del paso 4*)

Paso 16 Si $NFLAG = 0$ haga los pasos 17 y 18 (*Resultado rechazado.*)
también haga los pasos 19 y 20. (*Resultado aceptado.*)

Paso 17 Determine $h = h/2$. (*Valor nuevo para w rechazado, h disminuye*.)

Paso 18 Si $h < hmín$ entonces
SALIDA (‘*hmín excedida*’);
Determine $FLAG = 0$. (*Termina paso 16*)
(*Rama verdadera completada, el siguiente paso regresa al paso* 4.)

Paso 19 Determine $WO = y_1$; (*Valor de w nuevo aceptado*.)
$TO = TO + h$;
SALIDA (TO, WO, h).

Paso 20 Si $TO \geq b$ entonces determine $FLAG = 0$
(*Procedimiento completado con éxito.*)
también si $TO + h > b$ entonces determine $h = b - TO$
(*Termina en* $t = b$.)
también si ($k \leq 3$ y $h < 0.5(hmáx)$ entonces determine $h = 2h$.
(*Incrementa tamaño de paso, de ser posible*)
(*Fin del paso* 4 *y* 16)

Paso 21 PARE. ■

Ejemplo 1 Use el método de extrapolación con el tamaño de paso máximo *hmáx* = 0.2, tamaño de paso mínimo *hmín* = 0.01 y tolerancia *TOL* = 10^{-9} para aproximar la solución del problema de valor inicial

$$y' = y - t^2 + 1, \quad 0 \leq t \leq 2, \quad y(0) = 0.5.$$

Solución Para el primer paso del método de extrapolación, hacemos $w_0 = 0.5$, $t_0 = 0$, y $h = 0.2$. A continuación, calculamos

$$h_0 = h/2 = 0.1,$$
$$w_1 = w_0 + h_0 f(t_0, w_0) = 0.5 + 0.1(1.5) = 0.65,$$

y

$$w_2 = w_0 + 2h_0 f(t_0 + h_0, w_1) = 0.5 + 0.2(1.64) = 0.828,$$

y la primera aproximación para $y(0.2)$ es

$$y_{11} = \frac{1}{2}(w_2 + w_1 + h_0 f(t_0 + 2h_0, w_2)) = \frac{1}{2}(0.828 + 0.65 + 0.1 f(0.2, 0.828))$$
$$= 0.8284.$$

Para la segunda aproximación para $y(0.2)$, calculamos

$$h_1 = h/4 = 0.05,$$
$$w_1 = w_0 + h_1 f(t_0, w_0) = 0.5 + 0.05(1.5) = 0.575,$$
$$w_2 = w_0 + 2h_1 f(t_0 + h_1, w_1) = 0.5 + 0.1(1.5725) = 0.65725,$$
$$w_3 = w_1 + 2h_1 f(t_0 + 2h_1, w_2) = 0.575 + 0.1(1.64725) = 0.739725,$$

y

$$w_4 = w_2 + 2h_1 f(t_0 + 3h_1, w_3) = 0.65725 + 0.1(1.717225) = 0.8289725.$$

Entonces, la aproximación de la corrección del extremo es

$$y_{21} = \frac{1}{2}(w_4 + w_3 + h_1 f(t_0 + 4h_1, w_4))$$
$$= \frac{1}{2}(0.8289725 + 0.739725 + 0.05 f(0.2, 0.8289725)) = 0.8290730625.$$

Esto nos da la primera aproximación de extrapolación

$$y_{22} = y_{21} + \left(\frac{(1/4)^2}{(1/2)^2 - (1/4)^2}\right)(y_{21} - y_{11}) = 0.8292974167.$$

La tercera aproximación se encuentra al calcular

$$h_2 = h/6 = 0.0\overline{3},$$
$$w_1 = w_0 + h_2 f(t_0, w_0) = 0.55,$$
$$w_2 = w_0 + 2h_2 f(t_0 + h_2, w_1) = 0.6032592593,$$
$$w_3 = w_1 + 2h_2 f(t_0 + 2h_2, w_2) = 0.6565876543,$$
$$w_4 = w_2 + 2h_2 f(t_0 + 3h_2, w_3) = 0.7130317696,$$
$$w_5 = w_3 + 2h_2 f(t_0 + 4h_2, w_4) = 0.7696045871,$$
$$w_6 = w_4 + 2h_2 f(t_0 + 5h_2, w_5) = 0.8291535569,$$

y entonces la aproximación de corrección del extremo

$$y_{31} = \frac{1}{2}(w_6 + w_5 + h_2 f(t_0 + 6h_2, w_6) = 0.8291982979.$$

Ahora podemos encontrar dos aproximaciones extrapoladas

$$y_{32} = y_{31} + \left(\frac{(1/6)^2}{(1/4)^2 - (1/6)^2}\right)(y_{31} - y_{21}) = 0.8292984862$$

y

$$y_{33} = y_{32} + \left(\frac{(1/6)^2}{(1/2)^2 - (1/6)^2}\right)(y_{32} - y_{22}) = 0.8292986199.$$

Puesto que

$$|y_{33} - y_{22}| = 1.2 \times 10^{-6}$$

no satisface la tolerancia, necesitamos calcular por lo menos una fila de la tabla de extrapolación. Usamos $h_3 = h/8 = 0.025$ y calculamos w_1 con el método de Euler $w_2, \cdots, w_8$, a través del método de punto medio y aplicamos la corrección del extremo. Esto nos proporcionará la aproximación nueva y_{41}, que nos permite calcular la nueva fila de extrapolación

$$y_{41} = 0.8292421745 \quad y_{42} = 0.8292985873 \quad y_{43} = 0.8292986210 \quad y_{44} = 0.8292986211.$$

Al comparar $|y_{44} - y_{33}| = 1.2 \times 10^{-9}$, encontramos que la tolerancia de precisión no se ha alcanzado. Para obtener las entradas en la siguiente fila, usamos $h_4 = h/12 = 0.0\overline{6}$. Primero, calcule w_1 con el método de Euler, a continuación w_2 a través de w_{12} con el método de punto medio. Finalmente, utilice la corrección del extremo para obtener y_{51}. Las entradas restantes en la quinta columna se obtienen por medio de extrapolación y se muestran en la tabla 5.17. Puesto que $y_{55} - 0.8292986213$ está dentro de 10^{-9} de y_{44}, se acepta como la extrapolación para $y(0.2)$. El procedimiento provee una nueva aproximación $y(0.4)$. El conjunto completo de aproximaciones precisas para los lugares listados se proporciona en la tabla 5.18. ■

Tabla 5.17

$y_{1,1} = 0.8284000000$				
$y_{2,1} = 0.8290730625$	$y_{2,2} = 0.8292974167$			
$y_{3,1} = 0.8291982979$	$y_{3,2} = 0.8292984862$	$y_{3,3} = 0.8292986199$		
$y_{4,1} = 0.8292421745$	$y_{4,2} = 0.8292985873$	$y_{4,3} = 0.8292986210$	$y_{4,4} = 0.8292986211$	
$y_{5,1} = 0.8292735291$	$y_{5,2} = 0.8292986128$	$y_{5,3} = 0.8292986213$	$y_{5,4} = 0.8292986213$	$y_{5,5} = 0.8292986213$

Tabla 5.18

t_i	$y_i = y(t_i)$	w_i	h_i	k
0.200	0.8292986210	0.8292986213	0.200	5
0.400	1.2140876512	1.2140876510	0.200	4
0.600	1.6489405998	1.6489406000	0.200	4
0.700	1.8831236462	1.8831236460	0.100	5
0.800	2.1272295358	2.1272295360	0.100	4
0.900	2.3801984444	2.3801984450	0.100	7
0.925	2.4446908698	2.4446908710	0.025	8
0.950	2.5096451704	2.5096451700	0.025	3
1.000	2.6408590858	2.6408590860	0.050	3
1.100	2.9079169880	2.9079169880	0.100	7
1.200	3.1799415386	3.1799415380	0.100	6
1.300	3.4553516662	3.4553516610	0.100	8
1.400	3.7324000166	3.7324000100	0.100	5
1.450	3.8709427424	3.8709427340	0.050	7
1.475	3.9401071136	3.9401071050	0.025	3
1.525	4.0780532154	4.0780532060	0.050	4
1.575	4.2152541820	4.2152541820	0.050	3
1.675	4.4862274254	4.4862274160	0.100	4
1.775	4.7504844318	4.7504844210	0.100	4
1.825	4.8792274904	4.8792274790	0.050	3
1.875	5.0052154398	5.0052154290	0.050	3
1.925	5.1280506670	5.1280506570	0.050	4
1.975	5.2473151731	5.2473151660	0.050	8
2.000	5.3054719506	5.3054719440	0.025	3

La prueba de que el método presentado en el algoritmo 5.6 converge implica resultados a partir de la teoría de capacidad de suma; se puede encontrar en el artículo original de Gragg [Gr]. Existen otros procedimientos de extrapolación, algunos de los cuales utilizan técnicas de tamaño de paso variable. Para los procedimientos adicionales con base en los procesos de extrapolación, consulte los artículos de Bulirsch y Stoer [BS1], [BS2] y [BS3] o el texto de Stetter [Stet]. Los métodos usados por Bulirsch y Stoer implican interpolación con funciones racionales en lugar de la interpolación polinomial utilizada en el procedimiento de Gragg.

La sección Conjunto de ejercicios 5.8 está disponible en línea. Encuentre la ruta de acceso en las páginas preliminares.

5.9 Ecuaciones de orden superior y sistemas de ecuaciones diferenciales

Esta sección contiene una introducción a la solución numérica de problemas de valor inicial de orden superior. Las técnicas analizadas están limitadas a aquellas que transforman una ecuación de orden superior en un sistema de ecuaciones diferenciales de primer orden. Antes de analizar el procedimiento de transformación, se necesitan algunas observaciones respecto a los sistemas que implican ecuaciones diferenciales de primer orden.

Un **sistema de *m*-ésimo orden** de problemas de valor inicial de primer orden tiene la forma

$$\begin{aligned}
\frac{du_1}{dt} &= f_1(t, u_1, u_2, \ldots, u_m),\\
\frac{du_2}{dt} &= f_2(t, u_1, u_2, \ldots, u_m),\\
&\vdots\\
\frac{du_m}{dt} &= f_m(t, u_1, u_2, \ldots, u_m),
\end{aligned} \tag{5.45}$$

para $a \leq t \leq b$, con condiciones iniciales

$$u_1(a) = \alpha_1, \; u_2(a) = \alpha_2, \; \ldots, \; u_m(a) = \alpha_m. \tag{5.46}$$

El objetivo es encontrar m funciones $u_1(t), u_2(t), \ldots, u_m(t)$ que satisfacen cada una de las ecuaciones diferenciales junto con todas las condiciones iniciales.

Para analizar la existencia y unicidad de las soluciones para los sistemas de ecuaciones, necesitamos ampliar la definición de la condición de Lipschitz para las funciones de distintas variables.

Definición 5.16 Se dice que la función $f(t, y_1, \ldots, y_m)$, definida en el conjunto

$$D = \{(t, u_1, \ldots, u_m) \mid a \leq t \leq b \text{ y } -\infty < u_i < \infty, \text{ para cada } i = 1, 2, \ldots, m\},$$

satisface la **condición de Lipschitz** en D en las variables $u_1, u_2, \ldots, u_m$ si existe una constante $L > 0$ con

$$|f(t, u_1, \ldots, u_m) - f(t, z_1, \ldots, z_m)| \leq L \sum_{j=1}^{m} |u_j - z_j|, \tag{5.47}$$

para todas las $(t, u_1, \ldots, u_m)$ y $(t, z_1, \ldots, z_m)$ en D. ■

Al usar el teorema de valor medio se puede mostrar que si f y sus primeras derivadas parciales son continuas en D y si

$$\left| \frac{\partial f(t, u_1, \ldots, u_m)}{\partial u_i} \right| \leq L,$$

para cada $i = 1, 2, \ldots, m$ y todas las $(t, u_1, \ldots, u_m)$ en D, entonces f satisface la condición de Lipschitz en D con constante de Lipschitz L (consulte [BiR], p. 141). A continuación se muestra un teorema de existencia y unicidad básico, su demostración se puede encontrar en [BiR], p. 152-154.

Teorema 5.17 Suponga que

$$D = \{ (t, u_1, u_2, \ldots, u_m) \mid a \leq t \leq b \text{ y } -\infty < u_i < \infty, \text{ para cada } i = 1, 2, \ldots, m \},$$

y $f_i(t, u_1, \ldots, u_m)$, para cada $i = 1, 2, \ldots, m$ y que es continua y satisface la condición de Lipschitz en D. El sistema de ecuaciones diferenciales de primer orden (5.45), sujeto a las condiciones iniciales (5.46), tiene una única solución $u_1(t), \ldots, u_m(t)$, para $a \leq t \leq b$. ■

Los métodos para resolver sistemas de ecuaciones diferenciales de primer orden son generalizaciones de los métodos de ecuaciones de primer orden simples que se han presentado antes en este capítulo. Por ejemplo, el método clásico de Runge-Kutta de orden 4 dado por

$$\begin{aligned}
w_0 &= \alpha, \\
k_1 &= hf(t_i, w_i), \\
k_2 &= hf\left(t_i + \frac{h}{2}, w_i + \frac{1}{2}k_1\right), \\
k_3 &= hf\left(t_i + \frac{h}{2}, w_i + \frac{1}{2}k_2\right), \\
k_4 &= hf(t_{i+1}, w_i + k_3), \\
w_{i+1} &= w_i + \frac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4), \quad \text{para cada } i = 0, 1, \ldots, N-1,
\end{aligned}$$

utilizado para resolver el problema de valor inicial de primer orden

$$y' = f(t, y), \quad a \le t \le b, \quad y(a) = \alpha,$$

se generaliza de acuerdo con lo siguiente.

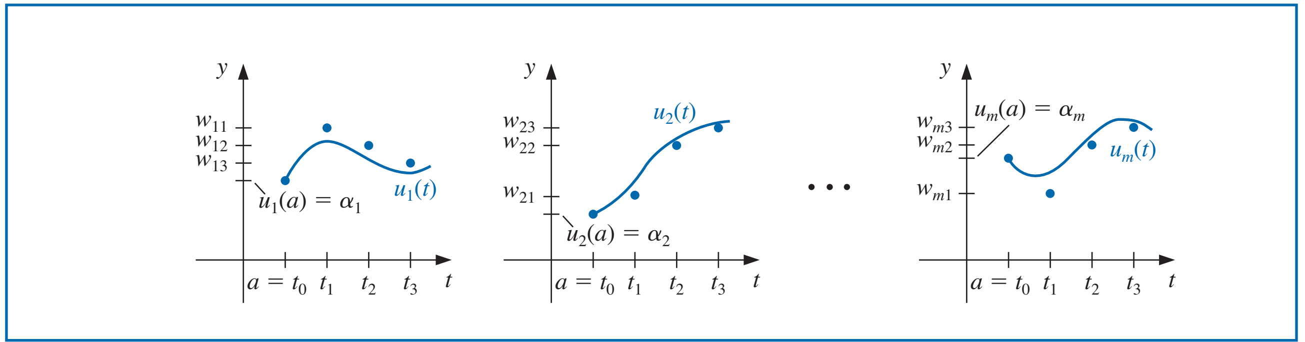
Si se selecciona un entero $N > 0$ y se establece $h = (b - a)/N$. La partición del intervalo $[a, b]$ en N subintervalos con los puntos de malla

$$t_j = a + jh, \quad \text{para cada } j = 0, 1, \ldots, N.$$

Use la notación w_{ij}, para cada $j = 0, 1, \ldots, N$ y $i = 1, 2, \ldots, m$, para denotar una aproximación $u_i(t_j)$. Es decir, w_{ij} aproxima la i-ésima solución $u_i(t)$ de (5.45) en el j-ésimo punto de malla t_j. Para las condiciones iniciales, establezca (consulte la figura 5.6)

$$w_{1,0} = \alpha_1, w_{2,0} = \alpha_2, \ldots, w_{m,0} = \alpha_m. \tag{5.48}$$

Figura 5.6



Suponga que los valores $w_{1,j}, w_{2,j}, \ldots, w_{m,j}$ se han calculado. Nosotros obtenemos $w_{1,j+1}, w_{2,j+1}, \ldots, w_{m,j+1}$ al calcular primero

$$k_{1,i} = hf_i(t_j, w_{1,j}, w_{2,j}, \ldots, w_{m,j}), \quad \text{para cada } i = 1, 2, \ldots, m, \tag{5.49}$$

$$k_{2,i} = hf_i\left(t_j + \frac{h}{2}, w_{1,j} + \frac{1}{2}k_{1,1}, w_{2,j} + \frac{1}{2}k_{1,2}, \ldots, w_{m,j} + \frac{1}{2}k_{1,m}\right), \tag{5.50}$$

para cada $i = 1, 2, \ldots, m$;

$$k_{3,i} = hf_i\left(t_j + \frac{h}{2}, w_{1,j} + \frac{1}{2}k_{2,1}, w_{2,j} + \frac{1}{2}k_{2,2}, \ldots, w_{m,j} + \frac{1}{2}k_{2,m}\right), \tag{5.51}$$

para cada $i = 1, 2, \ldots, m$;

$$k_{4,i} = hf_i(t_j + h, w_{1,j} + k_{3,1}, w_{2,j} + k_{3,2}, \ldots, w_{m,j} + k_{3,m}), \tag{5.52}$$

para cada $i = 1, 2, \ldots, m$; y entonces

$$w_{i,j+1} = w_{i,j} + \frac{1}{6}(k_{1,i} + 2k_{2,i} + 2k_{3,i} + k_{4,i}), \tag{5.53}$$

para cada $i = 1, 2, \ldots, m$. Observe que todos los valores $k_{1,1}, k_{1,2}, \ldots, k_{1,m}$ deben calcularse antes de que cualquiera de los términos de la forma $k_{2,i}$ se pueda determinar. En general, cada $k_{l,1}, k_{l,2}, \ldots, k_{l,m}$ se debe calcular antes que cualquiera de las expresiones $k_{l+1,i}$. El algoritmo 5.7 implementa el método Runge-Kutta de cuarto orden para sistemas de problemas de valor inicial.

ALGORITMO 5.7

Método Runge-Kutta para sistemas de ecuaciones diferenciales

Para aproximar la solución del sistema de m-ésimo orden de problemas de valor inicial de primer orden

$$u'_j = f_j(t, u_1, u_2, \ldots, u_m), \quad a \le t \le b, \quad \text{con} \quad u_j(a) = \alpha_j,$$

para $j = 1, 2, \ldots, m$ en $(N+1)$ números igualmente espaciados en el intervalo $[a, b]$:

ENTRADA extremos a, b; número de ecuaciones m; entero N; condiciones iniciales $\alpha_1, \ldots, \alpha_m$.

SALIDA aproximaciones w_j para $u_j(t)$ en $(N+1)$ valores de t.

Paso 1 Determine $h = (b-a)/N$;
$t = a$.

Paso 2 Para $j = 1, 2, \ldots, m$ determine $w_j = \alpha_j$.

Paso 3 SALIDA $(t, w_1, w_2, \ldots, w_m)$.

Paso 4 Para $i = 1, 2, \ldots, N$ haga los pasos 5–11.

Paso 5 Para $j = 1, 2, \ldots, m$ determine
$k_{1,j} = hf_j(t, w_1, w_2, \ldots, w_m)$.

Paso 6 Para $j = 1, 2, \ldots, m$ determine

$$k_{2,j} = hf_j\left(t + \tfrac{h}{2}, w_1 + \tfrac{1}{2}k_{1,1}, w_2 + \tfrac{1}{2}k_{1,2}, \ldots, w_m + \tfrac{1}{2}k_{1,m}\right).$$

Paso 7 Para $j = 1, 2, \ldots, m$ determine

$$k_{3,j} = hf_j\left(t + \tfrac{h}{2}, w_1 + \tfrac{1}{2}k_{2,1}, w_2 + \tfrac{1}{2}k_{2,2}, \ldots, w_m + \tfrac{1}{2}k_{2,m}\right).$$

Paso 8 Para $j = 1, 2, \ldots, m$ determine
$k_{4,j} = hf_j(t + h, w_1 + k_{3,1}, w_2 + k_{3,2}, \ldots, w_m + k_{3,m})$.

Paso 9 Para $j = 1, 2, \ldots, m$ determine
$w_j = w_j + (k_{1,j} + 2k_{2,j} + 2k_{3,j} + k_{4,j})/6$.

Paso 10 Determine $t = a + ih$.

Paso 11 SALIDA $(t, w_1, w_2, \ldots, w_m)$.

Paso 12 PARE. ■

Ilustración La ley de Kirchhoff establece que la suma de todos los cambios de voltaje alrededor de un circuito cerrado es cero. Esta ley implica que la corriente $I(t)$ en un circuito cerrado que contiene una resistencia de R ohms, una capacitancia de C faradios, una inductancia de L henrios y una fuente voltaje de $E(t)$ volts satisface la ecuación

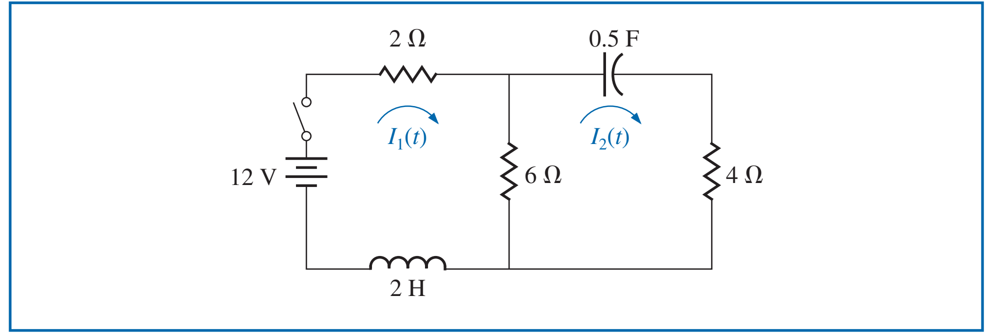
$$LI'(t) + RI(t) + \frac{1}{C}\int I(t)\,dt = E(t).$$

Las corrientes $I_1(t)$ y $I_2(t)$ en los ciclos izquierdo y derecho, respectivamente, del circuito mostrado en la figura 5.7 son las soluciones para el sistema de ecuaciones

$$2I_1(t) + 6[I_1(t) - I_2(t)] + 2I_1'(t) = 12,$$

$$\frac{1}{0.5}\int I_2(t)\,dt + 4I_2(t) + 6[I_2(t) - I_1(t)] = 0.$$

Figura 5.7



Si el interruptor en el circuito está cerrado en el tiempo $t = 0$, tenemos las condiciones iniciales $I_1(0) = 0$ y $I_2(0) = 0$. Resuelva para $I_1'(t)$ en la primera ecuación, al derivar la segunda ecuación y sustituyendo para $I_1'(t)$ se obtiene

$$I_1' = f_1(t, I_1, I_2) = -4I_1 + 3I_2 + 6, \quad I_1(0) = 0,$$

$$I_2' = f_2(t, I_1, I_2) = 0.6I_1' - 0.2I_2 = -2.4I_1 + 1.6I_2 + 3.6, \quad I_2(0) = 0.$$

La solución exacta para este sistema es

$$I_1(t) = -3.375e^{-2t} + 1.875e^{-0.4t} + 1.5,$$

$$I_2(t) = -2.25e^{-2t} + 2.25e^{-0.4t}.$$

Aplicaremos el método Runge-Kutta de orden 4 para este sistema con $h = 0.1$. Puesto que $w_{1,0} = I_1(0) = 0$ y $w_{2,0} = I_2(0) = 0$,

$$k_{1,1} = hf_1(t_0, w_{1,0}, w_{2,0}) = 0.1\ f_1(0, 0, 0) = 0.1\ (-4(0) + 3(0) + 6) = 0.6,$$

$$k_{1,2} = hf_2(t_0, w_{1,0}, w_{2,0}) = 0.1\ f_2(0, 0, 0) = 0.1\ (-2.4(0) + 1.6(0) + 3.6) = 0.36,$$

$$k_{2,1} = hf_1\left(t_0 + \frac{1}{2}h, w_{1,0} + \frac{1}{2}k_{1,1}, w_{2,0} + \frac{1}{2}k_{1,2}\right) = 0.1\ f_1(0.05, 0.3, 0.18)$$

$$= 0.1\ (-4(0.3) + 3(0.18) + 6) = 0.534,$$

$$k_{2,2} = hf_2\left(t_0 + \frac{1}{2}h, w_{1,0} + \frac{1}{2}k_{1,1}, w_{2,0} + \frac{1}{2}k_{1,2}\right) = 0.1\ f_2(0.05, 0.3, 0.18)$$

$$= 0.1\ (-2.4(0.3) + 1.6(0.18) + 3.6) = 0.3168.$$

Al generar las entradas restantes de manera similar obtenemos

$$k_{3,1} = (0.1)\,f_1(0.05, 0.267, 0.1584) = 0.54072,$$

$$k_{3,2} = (0.1)\,f_2(0.05, 0.267, 0.1584) = 0.321264,$$

$$k_{4,1} = (0.1)\,f_1(0.1, 0.54072, 0.321264) = 0.4800912,$$

$$k_{4,2} = (0.1)\,f_2(0.1, 0.54072, 0.321264) = 0.28162944.$$

Como consecuencia,

$$I_1(0.1) \approx w_{1,1} = w_{1,0} + \frac{1}{6}(k_{1,1} + 2k_{2,1} + 2k_{3,1} + k_{4,1})$$

$$= 0 + \frac{1}{6}\,(0.6 + 2(0.534) + 2(0.54072) + 0.4800912) = 0.5382552$$

y

$$I_2(0.1) \approx w_{2,1} = w_{2,0} + \frac{1}{6}(k_{1,2} + 2k_{2,2} + 2k_{3,2} + k_{4,2}) = 0.3196263.$$

Las entradas restantes en la tabla 5.19 se generan de manera similar. ■

Tabla 5.19

t_j	$w_{1,j}$	$w_{2,j}$	$\lvert I_1(t_j) - w_{1,j}\rvert$	$\lvert I_2(t_j) - w_{2,j}\rvert$
0.0	0	0	0	0
0.1	0.5382550	0.3196263	0.8285×10^{-5}	0.5803×10^{-5}
0.2	0.9684983	0.5687817	0.1514×10^{-4}	0.9596×10^{-5}
0.3	1.310717	0.7607328	0.1907×10^{-4}	0.1216×10^{-4}
0.4	1.581263	0.9063208	0.2098×10^{-4}	0.1311×10^{-4}
0.5	1.793505	1.014402	0.2193×10^{-4}	0.1240×10^{-4}

Ecuaciones diferenciales de orden superior

Muchos problemas físicos importantes, por ejemplo, circuitos eléctricos y sistemas de vibración, implican problemas de valor inicial cuyas ecuaciones tienen órdenes mayores que 1. No se requieren técnicas nuevas para resolverlos. Al reetiquetar las variables podemos reducir una ecuación diferencial de orden superior en un sistema de ecuaciones diferenciales de primer orden y, a continuación aplicar uno de los métodos que ya analizamos.

Un problema de valor inicial de m-ésimo orden general

$$y^{(m)}(t) = f(t, y, y', \ldots, y^{(m-1)}), \quad a \le t \le b,$$

con condiciones iniciales $y(a) = \alpha_1, y'(a) = \alpha_2, \ldots, y^{(m-1)}(a) = \alpha_m$, se pueden convertir en un sistema de ecuaciones de la forma (5.45) y (5.46).

Si $u_1(t) = y(t), u_2(t) = y'(t), \ldots,$ y $u_m(t) = y^{(m-1)}(t)$. Esto produce el sistema de primer orden

$$\frac{du_1}{dt} = \frac{dy}{dt} = u_2, \quad \frac{du_2}{dt} = \frac{dy'}{dt} = u_3, \quad \cdots, \quad \frac{du_{m-1}}{dt} = \frac{dy^{(m-2)}}{dt} = u_m,$$

y

$$\frac{du_m}{dt} = \frac{dy^{(m-1)}}{dt} = y^{(m)} = f(t, y, y', \ldots, y^{(m-1)}) = f(t, u_1, u_2, \ldots, u_m),$$

con condiciones iniciales

$$u_1(a) = y(a) = \alpha_1, \quad u_2(a) = y'(a) = \alpha_2, \quad \ldots, \quad u_m(a) = y^{(m-1)}(a) = \alpha_m.$$

Ejemplo 1 Transforme el problema de valor inicial de segundo orden

$$y'' - 2y' + 2y = e^{2t} \operatorname{sen} t, \qquad \text{para } 0 \le t \le 1, \quad \text{con } y(0) = -0.4, \; y'(0) = -0.6$$

en un sistema de problemas de valor inicial de primer orden y use el método Runge-Kutta con $h = 0.1$ para aproximar la solución.

Solución Sea $u_1(t) = y(t)$ y $u_2(t) = y'(t)$. Esto transforma la ecuación de segundo orden en el sistema

$$u_1'(t) = u_2(t),$$
$$u_2'(t) = e^{2t} \operatorname{sen} t - 2u_1(t) + 2u_2(t),$$

con condiciones iniciales $u_1(0) = -0.4, \; u_2(0) = -0.6$.

Las condiciones iniciales dan $w_{1,0} = -0.4$ y $w_{2,0} = -0.6$. Las ecuaciones (5.49) a (5.52) en la página 249 con $j = 0$ proporcionan

$$k_{1,1} = hf_1(t_0, w_{1,0}, w_{2,0}) = hw_{2,0} = -0.06,$$

$$k_{1,2} = hf_2(t_0, w_{1,0}, w_{2,0}) = h\left[e^{2t_0} \operatorname{sen} t_0 - 2w_{1,0} + 2w_{2,0}\right] = -0.04,$$

$$k_{2,1} = hf_1\left(t_0 + \frac{h}{2}, w_{1,0} + \frac{1}{2}k_{1,1}, w_{2,0} + \frac{1}{2}k_{1,2}\right) = h\left[w_{2,0} + \frac{1}{2}k_{1,2}\right] = -0.062,$$

$$k_{2,2} = hf_2\left(t_0 + \frac{h}{2}, w_{1,0} + \frac{1}{2}k_{1,1}, w_{2,0} + \frac{1}{2}k_{1,2}\right)$$

$$= h\left[e^{2(t_0+0.05)} \operatorname{sen}(t_0 + 0.05) - 2\left(w_{1,0} + \frac{1}{2}k_{1,1}\right) + 2\left(w_{2,0} + \frac{1}{2}k_{1,2}\right)\right]$$

$$= -0.03247644757,$$

$$k_{3,1} = h\left[w_{2,0} + \frac{1}{2}k_{2,2}\right] = -0.06162832238,$$

$$k_{3,2} = h\left[e^{2(t_0+0.05)} \operatorname{sen}(t_0 + 0.05) - 2\left(w_{1,0} + \frac{1}{2}k_{2,1}\right) + 2\left(w_{2,0} + \frac{1}{2}k_{2,2}\right)\right]$$

$$= -0.03152409237,$$

$$k_{4,1} = h\left[w_{2,0} + k_{3,2}\right] = -0.06315240924,$$

y

$$k_{4,2} = h\left[e^{2(t_0+0.1)} \operatorname{sen}(t_0 + 0.1) - 2(w_{1,0} + k_{3,1}) + 2(w_{2,0} + k_{3,2})\right] = -0.02178637298.$$

Por lo que,

$$w_{1,1} = w_{1,0} + \frac{1}{6}(k_{1,1} + 2k_{2,1} + 2k_{3,1} + k_{4,1}) = -0.4617333423$$

y

$$w_{2,1} = w_{2,0} + \frac{1}{6}(k_{1,2} + 2k_{2,2} + 2k_{3,2} + k_{4,2}) = -0.6316312421.$$

El valor w,$_{1,1}$ aproxima $u_1(0.1) = y(0.1) = 0.2e^{2(0.1)}(\operatorname{sen} 0.1 - 2\cos 0.1)$, y $w_{2,1}$ aproxima $u_2(0.1) = y'(0.1) = 0.2e^{2(0.1)}(4 \operatorname{sen} 0.1 - 3\cos 0.1)$.

El conjunto de valores $w_{1,j}$ y $w_{2,j}$, para $j = 0, 1, \ldots, 10$, se presenta en la tabla 5.20 y se comparan con los valores reales de $u_1(t) = 0.2e^{2t}(\operatorname{sen} t - 2\cos t)$ y $u_2(t) = u_1'(t) = 0.2e^{2t}(4 \operatorname{sen} t - 3\cos t)$. ■

Tabla 5.20

t_j	$y(t_j) = u_1(t_j)$	$w_{1,j}$	$y'(t_j) = u_2(t_j)$	$w_{2,j}$	$\lvert y(t_j) - w_{1,j}\rvert$	$\lvert y'(t_j) - w_{2,j}\rvert$
0.0	−0.40000000	−0.40000000	−0.6000000	−0.60000000	0	0
0.1	−0.46173297	−0.46173334	−0.6316304	−0.63163124	3.7×10^{-7}	7.75×10^{-7}
0.2	−0.52555905	−0.52555988	−0.6401478	−0.64014895	8.3×10^{-7}	1.01×10^{-6}
0.3	−0.58860005	−0.58860144	−0.6136630	−0.61366381	1.39×10^{-6}	8.34×10^{-7}
0.4	−0.64661028	−0.64661231	−0.5365821	−0.53658203	2.03×10^{-6}	1.79×10^{-7}
0.5	−0.69356395	−0.69356666	−0.3887395	−0.38873810	2.71×10^{-6}	5.96×10^{-7}
0.6	−0.72114849	−0.72115190	−0.1443834	−0.14438087	3.41×10^{-6}	7.75×10^{-7}
0.7	−0.71814890	−0.71815295	0.2289917	0.22899702	4.05×10^{-6}	2.03×10^{-6}
0.8	−0.66970677	−0.66971133	0.7719815	0.77199180	4.56×10^{-6}	5.30×10^{-6}
0.9	−0.55643814	−0.55644290	1.534764	1.5347815	4.76×10^{-6}	9.54×10^{-6}
1.0	−0.35339436	−0.35339886	2.578741	2.5787663	4.50×10^{-6}	1.34×10^{-5}

Los otros métodos de un paso se pueden extender a los sistemas de manera similar. Cuando los métodos de control de error como el Runge-Kutta-Fehlberg son extendidos, a cada componente de la solución numérica $(w_{1j}, w_{2j}, \ldots, w_{mj})$ se debe examinar su precisión. Si cualquiera de los componentes no es suficientemente preciso, toda la solución numérica $(w_{1j}, w_{2j}, \ldots, w_{mj})$ se debe volver a calcular.

Los métodos multipasos y las técnicas de indicador-corrector también se pueden ampliar a los sistemas. De nuevo, si se utiliza control de error, cada componente debe ser preciso. La extensión de la técnica de extrapolación para los sistemas también se puede realizar, pero la notación se vuelve bastante implicada. Si este tema es de interés, consulte [HNW1].

Los teoremas de convergencia y estimación de error para sistemas son similares a los considerados en la sección 5.10 para las ecuaciones individuales, excepto que las cotas están dadas en términos de normas de vectores, un tema considerado en el capítulo 7 (una buena referencia para estos teoremas es [Ge1], p. 45–72).

La sección Conjunto de ejercicios 5.9 está disponible en línea. Encuentre la ruta de acceso en las páginas preliminares.

5.10 Estabilidad

En este capítulo se ha presentado una serie de métodos para aproximar la solución para un problema de valor inicial. A pesar de que existen otras técnicas, seleccionamos los métodos descritos aquí porque, en general, satisfacen tres criterios:

- Su desarrollo es suficientemente claro, por lo que se puede comprender cómo y por qué funcionan.
- Uno o más métodos proporcionarán resultados satisfactorios para muchos de los problemas encontrados por los estudiantes de ciencias e ingeniería.
- Muchas de las técnicas más avanzadas y complejas están basadas en una o varios de los procedimientos descritos aquí.

Métodos de un paso

En esta sección analizamos porqué se espera que estos métodos proporcionen resultados satisfactorios cuando métodos similares no lo hacen. Antes de empezar este análisis, necesitamos presentar dos definiciones relacionadas con la convergencia de los métodos de ecuación de diferencia de un paso para la solución de esa ecuación diferencial conforme el tamaño de peso disminuye.

Definición 5.18 Se dice que un método de ecuación de diferencia de un paso con error de truncamiento local $\tau_i(h)$ en el i-ésimo paso es **consistente** con la ecuación diferencial que aproxima si

$$\lim_{h\to 0} \max_{1\le i\le N} |\tau_i(h)| = 0. \qquad \blacksquare$$

Un método de un paso es consistente si la ecuación de diferencia para el método se enfoca en la ecuación diferencial, conforme el tamaño de paso tiende a cero.

Observe que esta definición es una definición *local* ya que, para cada uno de los valores $\tau_i(h)$, suponemos que la aproximación w_{i-1} y la solución exacta $y(t_{i-1})$ es la misma. Un medio más realista para analizar los efectos de reducir h es determinar el efecto global del método. Éste es el error máximo del método sobre todo el rango de la aproximación, al suponer solamente que el método da el resultado exacto en el valor inicial.

Definición 5.19 Se dice que un método de ecuación de diferencia de un paso es **convergente** respecto a la ecuación diferencial que aproxima si

$$\lim_{h\to 0} \max_{1\le i\le N} |w_i - y(t_i)| = 0,$$

Un método es convergente si la solución de la ecuación de diferencia se enfoca en la solución de la ecuación diferencial conforme el tamaño de paso tiende a cero.

donde $y(t_i)$ denota el valor exacto de la solución de la ecuación diferencial y w_i es la aproximación obtenida a partir del método de diferencia en el i-ésimo paso. ■

Ejemplo 1 Muestre que el método de Euler es convergente.

Solución Examine la desigualdad (5.10) en la página 202, en la fórmula de la cota de error para el método de Euler, nosotros observamos que bajo la hipótesis del teorema 5.9

$$\max_{1\leq i\leq N} |w_i - y(t_i)| \leq \frac{Mh}{2L}|e^{L(b-a)} - 1|.$$

Sin embargo, *M, L y b* son todas las constantes, y

$$\lim_{h\to 0} \max_{1\leq i\leq N} |w_i - y(t_i)| \leq \lim_{h\to 0} \frac{Mh}{2L} \left|e^{L(b-a)} - 1\right| = 0.$$

De modo que el método de Euler es convergente respecto a una ecuación diferencial que satisface las condiciones de esta definición. La razón de convergencia es $O(h)$. ■

Un método consistente de un paso tiene la propiedad de que la ecuación de diferencia para el método aproxima la ecuación diferencial cuando el tamaño de paso tiende a cero. Por lo que el error de truncamiento local de un método consistente se aproxima a cero conforme el tamaño de paso se aproxima a cero.

El otro tipo de cota de error del problema que existe al usar métodos de diferencia o soluciones aproximadas para las ecuaciones diferenciales es una consecuencia de no usar resultados exactos. En la práctica, ni las condiciones iniciales ni la aritmética que se realiza por consiguiente se representan de manera exacta debido al error de redondeo asociado con la aritmética de dígitos finitos. En la sección 5.2 observamos que esta consideración puede conducir a dificultades incluso para el método convergente de Euler.

Un método es estable cuando los resultados dependen continuamente de los datos iniciales.

Para analizar esta situación, por lo menos en parte, tratamos de determinar los métodos que son **estables** en el sentido de que cambios pequeños o perturbaciones en las condiciones iniciales producen proporcionalmente pequeños cambios en las aproximaciones subsiguientes.

El concepto de estabilidad de una ecuación de diferencia de un paso es de alguna forma análoga para la condición de una ecuación diferencial bien planteada, por lo que no es sorprendente que la condición de Lipschitz aparezca aquí, como lo hizo en el teorema correspondiente para ecuaciones diferenciales, el teorema 5.6 en la sección 5.1.

La parte i) del siguiente teorema se preocupa por la estabilidad de un método de un paso. La prueba de este resultado no es difícil y se considera en el ejercicio 1. La parte ii) del teorema 5.20 se preocupa por condiciones suficientes para que un método consistente sea convergente. La parte iii) justifica la observación realizada en la sección 5.5 sobre controlar el error global de un método al controlar su error de truncamiento local e implica que cuando el error de truncamiento local tiene una razón de convergencia $O(h^n)$, el error global tendrá la misma razón de convergencia. Las pruebas de las partes ii) y iii) son más difíciles que las de la parte i) y se pueden encontrar dentro del material presentado en [Ge1], p. 57-58.

Teorema 5.20 Suponga que el problema de valor inicial

$$y' = f(t, y), \quad a \leq t \leq b, \quad y(a) = \alpha,$$

se aproxima por medio de un método de diferencia de un paso de la forma

$$w_0 = \alpha,$$
$$w_{i+1} = w_i + h\phi(t_i, w_i, h).$$

También suponga que existe un número $h_0 > 0$ y que $\phi(t, w, h)$ es continua y que satisface la condición de Lipschitz en la variable *w* con constante *L* de Lipschitz en el conjunto

$$D = \{ (t, w, h) \mid a \leq t \leq b \text{ y } -\infty < w < \infty, 0 \leq h \leq h_0 \}.$$

Entonces

i) El método es estable;

ii) El método de diferencia es convergente si y sólo si es consistente, lo cual es equivalente a

$$\phi(t, y, 0) = f(t, y), \quad \text{para todas las } a \leq t \leq b;$$

iii) Si existe una función τ y, para cada $i = 1, 2, \ldots, N$, el error de truncamiento local $\tau_i(h)$ satisface $|\tau_i(h)| \leq \tau(h)$ siempre que $0 \leq h \leq h_0$, entonces

$$|y(t_i) - w_i| \leq \frac{\tau(h)}{L} e^{L(t_i - a)}.$$

■

Ejemplo 2 El método modificado de Euler está dado por $w_0 = \alpha$,

$$w_{i+1} = w_i + \frac{h}{2}\left[f(t_i, w_i) + f(t_{i+1}, w_i + hf(t_i, w_i))\right], \quad \text{para } i = 0, 1, \ldots, N-1.$$

Verifique que este método es estable al mostrar que satisface la hipótesis del teorema 5.20.

Solución Para este método

$$\phi(t, w, h) = \frac{1}{2}f(t, w) + \frac{1}{2}f(t + h, w + hf(t, w)).$$

Si f satisface la condición de Lipschitz en $\{(t, w) \mid a \leq t \leq b \text{ y } -\infty < w < \infty\}$ en la variable w con constante L, entonces, puesto que

$$\phi(t, w, h) - \phi(t, \overline{w}, h) = \frac{1}{2}f(t, w) + \frac{1}{2}f(t + h, w + hf(t, w)) - \frac{1}{2}f(t, \overline{w}) - \frac{1}{2}f(t + h, \overline{w} + hf(t, \overline{w})),$$

la condición de Lipschitz en f conduce a

$$\begin{aligned}
|\phi(t, w, h) - \phi(t, \overline{w}, h)| &\leq \frac{1}{2}L|w - \overline{w}| + \frac{1}{2}L\left|w + hf(t, w) - \overline{w} - hf(t, \overline{w})\right| \\
&\leq L|w - \overline{w}| + \frac{1}{2}L\left|hf(t, w) - hf(t, \overline{w})\right| \\
&\leq L|w - \overline{w}| + \frac{1}{2}hL^2|w - \overline{w}| \\
&= \left(L + \frac{1}{2}hL^2\right)|w - \overline{w}|.
\end{aligned}$$

Por lo tanto, ϕ satisface la condición de Lipschitz en w en el conjunto

$$\{(t, w, h) \mid a \leq t \leq b, -\infty < w < \infty, \text{ y } 0 \leq h \leq h_0\},$$

para cualquier $h_0 > 0$ con constante

$$L' = L + \frac{1}{2}h_0L^2.$$

Finalmente, si f es continua en $\{(t, w) \mid a \leq t \leq b, -\infty < w < \infty\}$, entonces ϕ es continua en

$$\{(t, w, h) \mid a \leq t \leq b, -\infty < w < \infty, \text{ y } 0 \leq h \leq h_0\},$$

por lo que el teorema 5.20 implica que el método modificado de Euler es estable. Al hacer $h = 0$, tenemos

$$\phi(t, w, 0) = \frac{1}{2} f(t, w) + \frac{1}{2} f(t + 0, w + 0 \cdot f(t, w)) = f(t, w),$$

por lo que la condición de consistencia expresada en el teorema 5.20, parte ii), se mantiene. Por lo tanto, el método es convergente. Además, hemos visto que para este método, el error de truncamiento local es $O(h^2)$, por lo que la convergencia del método modificado de Euler también es $O(h^2)$. ■

Métodos multipasos

Para métodos multipasos, los problemas relacionados con la consistencia, la convergencia y la estabilidad son compuestos debido al número de aproximaciones que implica cada paso. Mientras en los métodos de un paso, la aproximación w_{i+1} depende directamente sólo de la aproximación previa w_i, los métodos multipasos utilizan por lo menos dos de las aproximaciones previas y los métodos comunes utilizados implican más.

El método general multipasos para aproximar la solución del problema de valor inicial

$$y' = f(t, y), \quad a \leq t \leq b, \quad y(a) = \alpha, \tag{5.54}$$

tiene la forma

$$\begin{aligned} &w_0 = \alpha, \quad w_1 = \alpha_1, \quad \ldots, \; w_{m-1} = \alpha_{m-1}, \\ &w_{i+1} = a_{m-1} w_i + a_{m-2} w_{i-1} + \cdots + a_0 w_{i+1-m} + hF(t_i, h, w_{i+1}, w_i, \ldots, w_{i+1-m}), \end{aligned} \tag{5.55}$$

para cada $i = m - 1, m, \ldots, N - 1$, donde $a_0, a_1, \ldots, a_{m+1}$ son constantes, y como siempre, $h = (b - a)/N$ y $t_i = a + ih$.

El error de truncamiento local para un método multipasos expresado en esta forma es

$$\begin{aligned} \tau_{i+1}(h) = &\frac{y(t_{i+1}) - a_{m-1} y(t_i) - \cdots - a_0 y(t_{i+1-m})}{h} \\ &- F(t_i, h, y(t_{i+1}), y(t_i), \ldots, y(t_{i+1-m})), \end{aligned}$$

para cada $i = m - 1, m, \ldots, N - 1$. Como en los métodos de un paso, el error de truncamiento local mide la forma en la que la solución y para la ecuación diferencial no puede satisfacer la ecuación de diferencia.

Por el método Adams-Bashforth de cuatro pasos, hemos observado que

$$\tau_{i+1}(h) = \frac{251}{720} y^{(5)}(\mu_i) h^4, \quad \text{para algunos } \mu_i \in (t_{i-3}, t_{i+1}),$$

mientras que el método Adams–Moulton de tres pasos se tiene

$$\tau_{i+1}(h) = -\frac{19}{720} y^{(5)}(\mu_i) h^4, \quad \text{para algunas } \mu_i \in (t_{i-2}, t_{i+1}),$$

siempre y cuando, por supuesto, $y \in C^5[a, b]$.

A lo largo del análisis, se realizan dos suposiciones respecto a la función F:

- Si $f \equiv 0$ (es decir, si la ecuación diferencial es homogénea), entonces $F \equiv 0$ también.
- F satisface la condición de Lipschitz respecto a $\{w_j\}$ en el sentido de que existe una constante L, y, para cada par de secuencias $\{v_j\}_{j=0}^N$ y $\{\tilde{v}_j\}_{j=0}^N$ y para $i = m-1, m, \ldots, N-1$, tenemos

$$|F(t_i, h, v_{i+1}, \ldots, v_{i+1-m}) - F(t_i, h, \tilde{v}_{i+1}, \ldots, \tilde{v}_{i+1-m})| \leq L \sum_{j=0}^{m} |v_{i+1-j} - \tilde{v}_{i+1-j}|.$$

Los métodos Adams-Bashforth explícito y Adams-Moulton implícito satisfacen ambas condiciones, siempre que f satisfaga la condición de Lipschitz (consulte el ejercicio 2).

El concepto de convergencia para métodos multipasos es el mismo que para los métodos de un paso.

- Un método multipasos es **convergente** si la solución de la ecuación de diferencia se aproxima a la solución de la ecuación diferencial conforme el tamaño de paso se aproxima a cero. Esto significa que $\text{lím}_{h\to 0} \text{máx}_{0\leq i\leq N} \mid w_i - y(t_i)| = 0$.

Para la consistencia, sin embargo, se presenta una situación algo diferente. De nuevo, queremos que un método multipasos sea consistente siempre que la ecuación de diferencia se aproxime a la ecuación diferencial conforme el tamaño de paso se aproxima a cero; es decir, el error de truncamiento local se aproxima a cero en cada paso conforme el tamaño de paso se aproxima a cero. La condición adicional se presenta debido al número de valores iniciales requerido para los métodos multipasos. Puesto que, normalmente, sólo el primer valor inicial $w_0 = \alpha$, es exacto, necesitamos requerir que los errores en todos los valores iniciales $\{\alpha_i\}$ se aproximen a cero conforme el tamaño de paso se aproxima a cero. Por lo que,

$$\lim_{h\to 0} |\tau_i(h)| = 0, \quad \text{para toda } i = m, m+1, \ldots, N, \quad \text{y} \tag{5.56}$$

$$\lim_{h\to 0} |\alpha_i - y(t_i)| = 0, \quad \text{para toda } i = 1, 2, \ldots, m-1, \tag{5.57}$$

debe ser verdadero para que un método multipasos de la forma (5.55) sea **consistente**. Observe que la ecuación (5.57) implica que un método multipasos no será consistente a menos que el método de un paso que genera los valores iniciales también sea consistente.

El siguiente teorema para los métodos multipasos es similar al teorema 5.20, parte iii), y provee una relación entre el error de truncamiento local y el error global de un método multipasos. Este proporciona la justificación teórica para intentar controlar el error global al controlar el error de truncamiento local. La demostración de una forma ligeramente más general de este teorema se puede encontrar en [IK], p. 387-388.

Teorema 5.21 Suponga que se aproxima el problema de valor inicial

$$y' = f(t, y), \quad a \leq t \leq b, \quad y(a) = \alpha,$$

con un método indicador-corrector explícito de Adams con una ecuación indicadora de Adams-Bashforth de m pasos

$$w_{i+1} = w_i + h[b_{m-1} f(t_i, w_i) + \cdots + b_0 f(t_{i+1-m}, w_{i+1-m})],$$

con error de truncamiento local $\tau_{i+1}(h)$, y una ecuación correctora implícita de Adams-Moulton de $(m-1)$ pasos

$$w_{i+1} = w_i + h\left[\tilde{b}_{m-1} f(t_{i+1}, w_{i+1}) + \tilde{b}_{m-2} f(t_i, w_i) + \cdots + \tilde{b}_0 f(t_{i+2-m}, w_{i+2-m})\right],$$

con error de truncamiento local $\tilde{\tau}_{i+1}(h)$. Además, suponga que $f(t, y)$ y $f_y(t, y)$ son continuas en $D = \{ (t, y) \mid a \leq t \leq b$ y $-\infty < y < \infty \}$ y f_y está acotada. Entonces, el error de truncamiento local $\sigma_{i+1}(h)$ del método indicador corrector es

$$\sigma_{i+1}(h) = \tilde{\tau}_{i+1}(h) + \tau_{i+1}(h)\tilde{b}_{m-1}\frac{\partial f}{\partial y}(t_{i+1}, \theta_{i+1}),$$

donde θ_{i+1} es un número entre cero y $h\tau_{i+1}(h)$.

Además, existen constantes k_1 y k_2 tal que

$$|w_i - y(t_i)| \leq \left[\max_{0 \leq j \leq m-1} \left| w_j - y(t_j) \right| + k_1\sigma(h) \right] e^{k_2(t_i - a)},$$

donde $\sigma(h) = \text{máx}_{m \leq j \leq N} |\sigma_j(h)|$. ■

Antes de analizar las conexiones entre consistencia, convergencia y estabilidad para los métodos multipasos, necesitamos considerar con más detalle la ecuación de diferencia para un método multipasos. Al hacerlo, descubriremos la razón para seleccionar los métodos Adams como nuestros métodos multipasos estándar.

Relacionado con la ecuación de diferencia (5.55) provisto al inicio de este análisis,

$$w_0 = \alpha, \ w_1 = \alpha_1, \ \ldots, \ w_{m-1} = \alpha_{m-1},$$
$$w_{i+1} = a_{m-1}w_i + a_{m-2}w_{i-1} + \cdots + a_0 w_{i+1-m} + hF(t_i, h, w_{i+1}, w_i, \ldots, w_{i+1-m}),$$

es un polinomio, llamado **polinomio característico** del método, dado por

$$P(\lambda) = \lambda^m - a_{m-1}\lambda^{m-1} - a_{m-2}\lambda^{m-2} - \cdots - a_1\lambda - a_0. \tag{5.58}$$

La estabilidad de un método multipasos respecto al error de redondeo está dictada por las magnitudes de los ceros del polinomio característico. Para observar esto, considere aplicar el método multipasos estándar (5.55) al problema trivial de valor inicial

$$y' \equiv 0, \quad y(a) = \alpha, \quad \text{donde } \alpha \neq 0. \tag{5.59}$$

Este problema tiene solución exacta $y(t) \equiv \alpha$. Al examinar las ecuaciones (5.27) y (5.28) en la sección 5.6 (consulte las páginas 226 y 227), podemos observar que cualquier método multipasos producirá, en teoría, la solución exacta $w_n = \alpha$ para todas las n. La única desviación de la solución exacta se debe al error de redondeo del método.

El lado derecho de la ecuación diferencial en (5.59) tiene $f(t, y) \equiv 0$, por lo que mediante la suposición (1), tenemos $F(t_i, h, w_{i+1}, w_{i+2}, \ldots, w_{i+1-m}) = 0$ en la ecuación de diferencia (5.55). Como consecuencia, la forma estándar de esta ecuación se convierte en

$$w_{i+1} = a_{m-1}w_i + a_{m-2}w_{i-1} + \cdots + a_0 w_{i+1-m}. \tag{5.60}$$

Suponiendo que λ es uno de los ceros del polinomio característico relacionado con la ecuación (5.55). Entonces $w_n = \lambda^n$ para cada n es una solución para la ecuación (5.59) ya que

$$\lambda^{i+1} - a_{m-1}\lambda^i - a_{m-2}\lambda^{i-1} - \cdots - a_0\lambda^{i+1-m} = \lambda^{i+1-m}[\lambda^m - a_{m-1}\lambda^{m-1} - \cdots - a_0] = 0.$$

En efecto, si $\lambda_1, \lambda_2, \ldots, \lambda_m$ son raíces distintas del polinomio característico para la ecuación (5.55) se puede mostrar de que *todas* las soluciones de la ecuación (5.60) se pueden expresar en la forma

$$w_n = \sum_{i=1}^{m} c_i\lambda_i^n, \tag{5.61}$$

para algún conjunto único de constantes $c_1, c_2, \ldots, c_m$.

Puesto que la solución exacta para la ecuación (5.59) es $y(t) = \alpha$, la selección $w_n = \alpha$, para todas las n, es una solución para la ecuación (5.60). Usando este hecho en la ecuación (5.60) obtenemos

$$0 = \alpha - \alpha a_{m-1} - \alpha a_{m-2} - \cdots - \alpha a_0 = \alpha[1 - a_{m-1} - a_{m-2} - \cdots - a_0].$$

Esto implica que $\lambda = 1$ es uno de los ceros del polinomio característico (5.58). Supondremos que en la representación (5.61), esta solución está descrita por $\lambda_1 = 1$ y $c_1 = \alpha$, por lo que todas las soluciones de la ecuación (5.59) se expresan como

$$w_n = \alpha + \sum_{i=2}^{m} c_i \lambda_i^n. \tag{5.62}$$

Si todos los cálculos fueran exactos, todas las constantes $c_2, c_3, \ldots, c_m$ serían cero. En la práctica, las constantes $c_2, c_3, \ldots, c_m$ no son cero debido al error de redondeo. De hecho, el error de redondeo crece de manera exponencial a menos que $|\lambda_i| \leq 1$ para cada una de las raíces $\lambda_2, \lambda_3, \ldots, \lambda_m$. Mientras más pequeña sea la magnitud de estas raíces, más estable será el método respecto al crecimiento del error de redondeo.

Al derivar la ecuación (5.62) realizamos la suposición simplificadora de que los ceros del polinomio característico son distintos. La situación es similar cuando se presentan múltiples ceros. Por ejemplo, si $\lambda_k = \lambda_{k+1} = \cdots = \lambda_{k+p}$ para algunas k y p, simplemente requiere reemplazar la suma

$$c_k \lambda_k^n + c_{k+1} \lambda_{k+1}^n + \cdots + c_{k+p} \lambda_{k+p}^n$$

en (5.62) con

$$c_k \lambda_k^n + c_{k+1} n \lambda_k^{n-1} + c_{k+2} n(n-1) \lambda_k^{n-2} + \cdots + c_{k+p}[n(n-1)\cdots(n-p+1)] \lambda_k^{n-p}. \tag{5.63}$$

(Consulte [He2], p. 119–145). A pesar de que se modifica la forma de la solución, el error de redondeo $|\lambda_k| > 1$ sigue creciendo exponencialmente.

Aunque sólo hemos considerado el caso especial de aproximación de los problemas de valor inicial de la forma (5.59), las características de estabilidad para esta ecuación determinan la estabilidad de la situación cuando $f(t, y)$ no es idénticamente cero. Esto porque la solución para la ecuación homogénea (5.59) está integrada en la solución de cualquier ecuación. Las siguientes definiciones están motivadas por este análisis.

Definición 5.22 Si $\lambda_1, \lambda_2, \ldots, \lambda_m$ denota las raíces (no necesariamente distintas) de la ecuación característica

$$P(\lambda) = \lambda^m - a_{m-1}\lambda^{m-1} - \cdots - a_1\lambda - a_0 = 0$$

relacionada con el método de diferencias multipasos

$$w_0 = \alpha, \quad w_1 = \alpha_1, \quad \ldots, \quad w_{m-1} = \alpha_{m-1}$$

$$w_{i+1} = a_{m-1}w_i + a_{m-2}w_{i-1} + \cdots + a_0 w_{i+1-m} + hF(t_i, h, w_{i+1}, w_i, \ldots, w_{i+1-m}).$$

Si $|\lambda_i| \leq 1$, para cada $i = 1, 2, \ldots, m$, y todas las raíces con valor absoluto 1 son raíces simples, entonces se dice que el método de diferencia satisface la **condición de raíz**. ■

Definición 5.23

i) Los métodos que satisfacen la condición de raíz y tienen $\lambda = 1$ como la única raíz de la ecuación característica con magnitud uno reciben el nombre de **firmemente estables**.

ii) Los métodos que satisfacen la condición de raíz y tienen más de una raíz distinta con magnitud uno reciben el nombre de **débilmente estables**.

iii) Los métodos que no satisfacen la condición de raíz reciben el nombre de **inestables**.

■

La consistencia y convergencia de un método multipasos están relacionadas de cerca con la estabilidad del redondeo del método. El siguiente teorema detalla estas conexiones. Para la demostración de este resultado y la teoría en la que está basado, consulte [IK] p. 410–417

Definición 5.24 Un método multipasos de la forma

$$w_0 = \alpha, \quad w_1 = \alpha_1, \quad \ldots, \quad w_{m-1} = \alpha_{m-1},$$

$$w_{i+1} = a_{m-1}w_i + a_{m-2}w_{i-1} + \cdots + a_0 w_{i+1-m} + hF(t_i, h, w_{i+1}, w_i, \ldots, w_{i+1-m})$$

es estable si y sólo si satisface la condición raíz. Además, si el método de diferencia es consistente con la ecuación diferencial, entonces el método es estable si y sólo si es convergente. ■

Ejemplo 3 El método Adams–Bashforth de cuarto orden se puede expresar como

$$w_{i+1} = w_i + hF(t_i, h, w_{i+1}, w_i, \ldots, w_{i-3}),$$

donde

$$F(t_i, h, w_{i+1}, , \ldots, w_{i-3}) = \frac{h}{24}[55f(t_i, w_i) - 59f(t_{i-1}, w_{i-1}) + 37f(t_{i-2}, w_{i-2}) - 9f(t_{i-3}, w_{i-3})].$$

Muestre que este método es firmemente estable.

Solución En este caso, tenemos $m = 4$, $a_0 = 0$, $a_1 = 0$, $a_2 = 0$ y $a_3 = 1$, por lo que la ecuación característica para este método Adams-Bashforth es

$$0 = P(\lambda) = \lambda^4 - \lambda^3 = \lambda^3(\lambda - 1).$$

Este polinomio tiene raíces $\lambda_1 = 1$, $\lambda_2 = 0$, $\lambda_3 = 0$ y $\lambda_4 = 0$. Por lo tanto, satisface la condición de raíz y es firmemente estable.

El método Adams–Moulton tiene un polinomio característico similar, $P(\lambda) = \lambda^3 - \lambda^2$, con ceros $\lambda_1 = 1$, $\lambda_2 = 0$ y $\lambda_3 = 0$, y también es firmemente estable. ■

Ejemplo 4 Muestre que el método Milne de cuarto orden, el método explícito multipasos provisto por

$$w_{i+1} = w_{i-3} + \frac{4h}{3}\left[2f(t_i, w_i) - f(t_{i-1}, w_{i-1}) + 2f(t_{i-2}, w_{i-2})\right],$$

satisface la condición raíz, pero sólo es débilmente estable.

Solución La ecuación característica para este método $0 = P(\lambda) = \lambda^4 - 1$, tiene cuatro raíces con magnitud 1: $\lambda_1 = 1$, $\lambda_2 = -1$, $\lambda_3 = i$, y $\lambda_4 = -i$. Puesto que todas las raíces tienen magnitud 1, el método satisface la condición de raíz. Sin embargo, existen raíces múltiples con magnitud 1, por lo que el método es débilmente estable. ■

Ejemplo 5 Aplique el método Adams-Bashforth de cuarto orden firmemente estable y el método Milne débilmente estable con $h = 0.1$ para el problema de valor inicial

$$y' = -6y + 6, \quad 0 \le t \le 1, \quad y(0) = 2,$$

que tiene la solución exacta $y(t) = 1 + e^{-6t}$.

Solución Los resultados en la tabla 5.21 muestran los efectos de un método débilmente estable *versus* uno firmemente estable para este problema. ■

Tabla 5.21

t_i	Exacto $y(t_i)$	Método Adams–Bashforth w_i	Error $\lvert y_i - w_i \rvert$	Método de Milne w_i	Error $\lvert y_i - w_i \rvert$
0.10000000		1.5488116		1.5488116	
0.20000000		1.3011942		1.3011942	
0.30000000		1.1652989		1.1652989	
0.40000000	1.0907180	1.0996236	8.906×10^{-3}	1.0983785	7.661×10^{-3}
0.50000000	1.0497871	1.0513350	1.548×10^{-3}	1.0417344	8.053×10^{-3}
0.60000000	1.0273237	1.0425614	1.524×10^{-2}	1.0486438	2.132×10^{-2}
0.70000000	1.0149956	1.0047990	1.020×10^{-2}	0.9634506	5.154×10^{-2}
0.80000000	1.0082297	1.0359090	2.768×10^{-2}	1.1289977	1.208×10^{-1}
0.90000000	1.0045166	0.9657936	3.872×10^{-2}	0.7282684	2.762×10^{-1}
1.00000000	1.0024788	1.0709304	6.845×10^{-2}	1.6450917	6.426×10^{-1}

La razón para seleccionar el método Adams–Bashforth-Moulton como nuestra técnica indicador-corrector de cuarto orden en la sección 5.6 sobre el método del Milne-Simpson del mismo orden es que tanto el método Adams-Bashforth como el Adams-Moulton son firmemente estables. Estos tienen más posibilidades de proporcionar aproximaciones precisas para una clase más amplia de problemas que el método indicador-corrector con base en las técnicas de Milne y Simpson, ambos débilmente estables.

La sección Conjunto de ejercicios está disponible en línea. Encuentre la ruta de acceso en las páginas preliminares.



5.11 Ecuaciones diferenciales rígidas

Todos los métodos para aproximar la solución de los problemas de valor inicial tienen términos de error que implican una derivada superior de la solución de la ecuación. Si se puede acotar de manera razonable la derivada, entonces el método tendrá una cota de error previsible que se puede utilizar para calcular la precisión de la aproximación. Incluso si la derivada aumenta conforme los pasos aumentan, el error se puede mantener en control relativo, siempre y cuando la solución también crezca en magnitud. Sin embargo, con frecuencia surgen problemas cuando la magnitud de la derivada aumenta, pero la solución no lo hace. En esta situación, el error puede aumentar tanto que domina los cálculos. Los problemas de valor inicial para los que es probable que esto se presente reciben el nombre de **ecuaciones rígidas** y son bastante comunes, de modo especial en el estudio de vibraciones, reacciones químicas y circuitos eléctricos.

Los sistemas rígidos derivan su nombre del movimiento de sistemas de masa-resorte que tienen grandes constantes de resorte.

Las ecuaciones diferenciales rígidas se caracterizan como aquellas cuya solución exacta tiene un término de la forma e^{-ct}, donde c es una constante positiva grande. Normalmente, esto sólo es una parte de la solución, llamada solución *transitoria*. La parte más importante de la solución se llama *solución de estado estable*. La parte transitoria de una ecuación rígida tiende rápidamente a cero conforme t aumenta, pero debido a que la enésima derivada de este término tiene magnitud $c^n e^{-ct}$, la derivada no decae tan rápido. De hecho, puesto que la derivada en el término de error se evalúa no en t sino en un número entre cero y t, los términos derivados pueden incrementar conforme t aumenta, y por cierto, de manera muy rápida. Afortunadamente, en general, las ecuaciones rígidas se pueden predecir a partir de un problema físico, desde el que se derivan las ecuaciones, y, con cuidado, el error se puede mantener bajo control. La manera en la que esto se hace se considera en esta sección.

Ilustración El sistema de problemas de valor inicial

$$u_1' = 9u_1 + 24u_2 + 5\cos t - \frac{1}{3}\operatorname{sen} t, \quad u_1(0) = \frac{4}{3}$$
$$u_2' = -24u_1 - 51u_2 - 9\cos t + \frac{1}{3}\operatorname{sen} t, \quad u_2(0) = \frac{2}{3}$$

tiene la única solución

$$u_1(t) = 2e^{-3t} - e^{-39t} + \frac{1}{3}\cos t, \quad u_2(t) = -e^{-3t} + 2e^{-39t} - \frac{1}{3}\cos t.$$

El término transitorio e^{-39t} en la solución causa que este sistema sea rígido. Al aplicar el algoritmo 5.7, el método Runge-Kutta de cuarto orden para sistemas, da los resultados listados en la tabla 5.22. Cuando $h = 0.05$, los resultados de estabilidad y las aproximaciones son precisas. Al incrementar el tamaño de paso a $h = 0.1$, sin embargo, conduce a los resultados desastrosos mostrados en la tabla. ■

Tabla 5.22

t	$u_1(t)$	$w_1(t)$ $h = 0.05$	$w_1(t)$ $h = 0.1$	$u_2(t)$	$w_2(t)$ $h = 0.05$	$w_2(t)$ $h = 0.1$
0.1	1.793061	1.712219	−2.645169	−1.032001	−0.8703152	7.844527
0.2	1.423901	1.414070	−18.45158	−0.8746809	−0.8550148	38.87631
0.3	1.131575	1.130523	−87.47221	−0.7249984	−0.7228910	176.4828
0.4	0.9094086	0.9092763	−934.0722	−0.6082141	−0.6079475	789.3540
0.5	0.7387877	9.7387506	−1760.016	−0.5156575	−0.5155810	3520.00
0.6	0.6057094	0.6056833	−7848.550	−0.4404108	−0.4403558	15697.84
0.7	0.4998603	0.4998361	−34989.63	−0.3774038	−0.3773540	69979.87
0.8	0.4136714	0.4136490	−155979.4	−0.3229535	−0.3229078	311959.5
0.9	0.3416143	0.3415939	−695332.0	−0.2744088	−0.2743673	1390664.
1.0	0.2796748	0.2796568	−3099671.	−0.2298877	−0.2298511	6199352.

Aunque a menudo la rigidez se relaciona con los sistemas de ecuaciones diferenciales, las características de la aproximación de un método numérico particular aplicado a un sistema rígido se pueden predecir al examinar el error producido cuando se aplica el método a una *ecuación de prueba* simple,

$$y' = \lambda y, \quad y(0) = \alpha, \quad \text{donde } \lambda < 0. \tag{5.64}$$

La solución para esta ecuación es $y(t) = \alpha e^{\lambda t}$, que contiene la solución transitoria $e^{\lambda t}$. La solución de estado estable es cero, por lo que las características de un método son fáciles de determinar. (Un análisis más completo del error de redondeo relacionado con sistemas rígidos requiere examinar la ecuación de prueba cuando λ es un número complejo con parte real negativa; consulte [Ge1], p. 222.)

Primero, considere el método de Euler aplicado a la ecuación de prueba. Si $h = (b - a)/N$ y $t_j = jh$, para $j = 0, 1, 2, \ldots, N$, la ecuación (5.8) en la página 266 implica que

$$w_0 = \alpha, \quad \text{y} \quad w_{j+1} = w_j + h(\lambda w_j) = (1 + h\lambda)w_j,$$

por lo que

$$w_{j+1} = (1 + h\lambda)^{j+1} w_0 = (1 + h\lambda)^{j+1}\alpha, \quad \text{para } j = 0, 1, \ldots, N - 1. \tag{5.65}$$

Puesto que la solución exacta es $y(t) = \alpha e^{\lambda t}$, el error absoluto es

$$|y(t_j) - w_j| = \left|e^{jh\lambda} - (1 + h\lambda)^j\right| |\alpha| = \left|(e^{h\lambda})^j - (1 + h\lambda)^j\right| \; |\alpha|,$$

y la precisión se determina por qué tan bien el término $1+h\lambda$ aproxima $e^{h\lambda}$. Cuando $\lambda < 0$, la solución exacta $(e^{h\lambda})^j$ tiende a cero conforme j aumenta, pero con la ecuación (5.65), la aproximación tendrá esta propiedad sólo si $|1 + h\lambda| < 1$, lo cual implica que $-2 < h\lambda < 0$. Esto restringe efectivamente el tamaño de paso h para el método de Euler para satisfacer $h < 2/|\lambda|$.

Ahora, suponga que se introduce un error de redondeo δ_0 en la condición inicial para el método de Euler,

$$w_0 = \alpha + \delta_0.$$

En el j-ésimo paso, el error de redondeo es

$$\delta_j = (1 + h\lambda)^j \delta_0.$$

Puesto que $\lambda < 0$, la condición para el control del crecimiento del error de redondeo es la misma que la condición para controlar el error absoluto $|1+h\lambda| < 1$, que implica que $h < 2/|\lambda|$. Por lo que

- Se espera que el método de Euler sea estable para

$$y' = \lambda y, \quad y(0) = \alpha, \quad \text{en donde } \lambda < 0,$$

sólo si el tamaño de paso h es menor a $2/|\lambda|$.

La situación es similar para otros métodos de un paso. En general, existe una función Q con la propiedad de que el método de diferencia, cuando se aplica a la ecuación de prueba, da

$$w_{i+1} = Q(h\lambda)w_i. \tag{5.66}$$

La precisión del método depende de qué tan bien $Q(h\lambda)$ aproxima $e^{h\lambda}$, y el error crecerá sin cota si $|Q(h\lambda)| > 1$. El método de Taylor de enésimo orden, por ejemplo, tendrá estabilidad respecto tanto al crecimiento del error de redondeo como al error absoluto, siempre y cuando se seleccione h para satisfacer

$$\left|1 + h\lambda + \frac{1}{2}h^2\lambda^2 + \cdots + \frac{1}{n!}h^n\lambda^n\right| < 1.$$

El ejercicio 10 examina el caso específico cuando el método es el clásico Runge-Kutta de cuarto orden, que es, fundamentalmente, el método de Taylor de cuarto orden.

Cuando se aplica un método multipasos de la forma (5.54) a la ecuación de prueba, el resultado es

$$w_{j+1} = a_{m-1}w_j + \cdots + a_0 w_{j+1-m} + h\lambda(b_m w_{j+1} + b_{m-1}w_j + \cdots + b_0 w_{j+1-m}),$$

para $j = m-1, \ldots, N-1$, o

$$(1 - h\lambda b_m)w_{j+1} - (a_{m-1} + h\lambda b_{m-1})w_j - \cdots - (a_0 + h\lambda b_0)w_{j+1-m} = 0.$$

Relacionado con esta ecuación de diferencia homogénea se encuentra el **polinomio característico**

$$Q(z, h\lambda) = (1 - h\lambda b_m)z^m - (a_{m-1} + h\lambda b_{m-1})z^{m-1} - \cdots - (a_0 + h\lambda b_0).$$

Este polinomio es similar al polinomio característico (5.58), pero también incorpora la ecuación de prueba. La teoría aquí iguala el análisis de estabilidad en la sección 5.10.

Suponga que $w_0, \ldots, w_{m-1}$ están determinados y, para $h\lambda$, sean $\beta_1, \ldots, \beta_m$ los ceros del polinomio $Q(z, h\lambda)$. Si $\beta_1, \ldots, \beta_m$ son distintos, entonces existe $c_1, \ldots, c_m$ con

$$w_j = \sum_{k=1}^{m} c_k(\beta_k)^j, \quad \text{para } j = 0, \ldots, N. \tag{5.67}$$

Si $Q(z, h\lambda)$ tiene múltiples ceros, w_j se define de manera similar. (Consulte la ecuación (5.63) en la sección 5.10.) Si w_j aproxima con precisión $y(t_j) = e^{jh\lambda} = (e^{h\lambda})^j$, entonces todos los ceros β_k deben satisfacer $|\beta_k| < 1$; de lo contrario, ciertas opciones de α resultarán en $c_k \neq 0$, y el término $c_k(\beta_k)^j$ no decaerá a cero.

Ilustración La ecuación diferencial de prueba

$$y' = -30y, \quad 0 \leq t \leq 1.5, \quad y(0) = \frac{1}{3}$$

tiene la solución exacta $y = \frac{1}{3}e^{-30t}$. Al utilizar $h = 0.1$ para el algoritmo 5.1 de Euler, el algoritmo 5.2 de Runge-Kutta de cuarto orden y el algoritmo 5.4 de Adams indicador-corrector da los resultados en $t = 1.5$ en la tabla 5.23. ■

Tabla 5.23

Solución exacta	9.54173×10^{-21}
Método de Euler	-1.09225×10^{4}
Método Runge-Kutta	3.95730×10^{1}
Método indicador corrector	8.03840×10^{5}

Las imprecisiones en la ilustración se deben al hecho de que $|Q(h\lambda)| > 1$ para el método de Euler y el método Runge-Kutta y que $Q(z, h\lambda)$ tiene ceros con módulos que exceden a uno para el método indicador-corrector. Para aplicar estos métodos a este problema, se debe reducir el tamaño de paso. La siguiente definición se usa para describir la cantidad de reducción de tamaño de paso requerida

Definición 5.25 La **región R de estabilidad absoluta** para un método de un paso es $R = \{h\lambda \in \mathcal{C} \mid |Q(h\lambda)| < 1\}$, y para un método multipasos es $R = \{h\lambda \in \mathcal{C} \mid |\beta_k| < 1$, para todos los ceros β_k de $Q(z, h\lambda)\}$. ■

Las ecuaciones (5.66) y (5.67) implican que se puede aplicar un método de manera efectiva a una ecuación rígida sólo si $h\lambda$ se encuentra en la región de estabilidad absoluta del método, lo cual, para un problema determinado, establece una restricción en el tamaño de h. Aunque el término exponencial en la solución exacta decae rápidamente a cero, λh debe permanecer dentro de la región de estabilidad absoluta a lo largo del intervalo de t valores para que la aproximación tienda a cero y el crecimiento del error esté bajo control. Esto significa que, a pesar de que h se podría incrementar normalmente debido a las consideraciones del error de truncamiento, el criterio absoluto de estabilidad fuerza h a continuar siendo pequeño. Los métodos de tamaño de paso variable son especialmente vulnerables a este problema ya que una revisión del error de truncamiento local podría indicar que el tamaño de paso puede aumentar. Esto podría resultar inadvertidamente en $h\lambda$ fuera de la región de estabilidad absoluta.

En general, la región de estabilidad absoluta de un método es el factor crítico al producir aproximaciones precisas para sistemas rígidos, por lo que los métodos numéricos se buscan con una región de estabilidad absoluta tan grande como sea posible. Se dice que un método numérico es ***A*-estable** si su región R de estabilidad absoluta contiene todo el plano medio izquierdo.

Este método es implícito porque involucra w_{j+1} en ambos lados de la ecuación.

El **método trapezoidal implícito**, determinado por

$$w_0 = \alpha, \tag{5.68}$$

$$w_{j+1} = w_j + \frac{h}{2}\left[f(t_{j+1}, w_{j+1}) + f(t_j, w_j)\right], \quad 0 \leq j \leq N-1,$$

es un método A-estable (consulte el ejercicio 14) y es el único método multipasos A-estable. A pesar de que el método trapezoidal no brinda aproximaciones precisas para los tamaños de paso grandes, su error no aumentará de manera exponencial.

Las técnicas que se usan con mayor frecuencia para los sistemas rígidos son métodos multipasos implícitos. En general, w_{i+1} se obtiene al resolver una ecuación no lineal o un sistema no lineal de manera iterativa, a menudo con el método de Newton. Considere, por ejemplo, el método trapezoidal implícito

$$w_{j+1} = w_j + \frac{h}{2}[f(t_{j+1}, w_{j+1}) + f(t_j, w_j)].$$

Al calcular t_j, t_{j+1}, y w_j, necesitamos determinar w_{i+1}, la solución para

$$F(w) = w - w_j - \frac{h}{2}[f(t_{j+1}, w) + f(t_j, w_j)] = 0. \tag{5.69}$$

Para aproximar esta solución, seleccione $w_{j+1}^{(0)}$, normalmente como w_j, y genere $w_{j+1}^{(k)}$ al aplicar el método de Newton a la ecuación (5.69)

$$\begin{aligned} w_{j+1}^{(k)} &= w_{j+1}^{(k-1)} - \frac{F\left(w_{j+1}^{(k-1)}\right)}{F'\left(w_{j+1}^{(k-1)}\right)} \\ &= w_{j+1}^{(k-1)} - \frac{w_{j+1}^{(k-1)} - w_j - \frac{h}{2}\left[f(t_j, w_j) + f\left(t_{j+1}, w_{j+1}^{(k-1)}\right)\right]}{1 - \frac{h}{2} f_y\left(t_{j+1}, w_{j+1}^{(k-1)}\right)} \end{aligned}$$

hasta que $|w_{j+1}^{(k)} - w_{j+1}^{(k-1)}|$ sea suficientemente pequeña. Éste es el procedimiento que se aplica en el algoritmo 5.8. Normalmente, sólo se requieren tres o cuatro operaciones por paso debido a la convergencia cuadrática del método de Newton.

El método de la secante se puede utilizar como una alternativa para el método de Newton en la ecuación (5.69) pero entonces, se requieren dos aproximaciones iniciales diferentes para w_{j+1}. Para usar el método de la secante, la práctica usual es hacer $w_{j+1}^{(0)} = w_j$ y obtener $w_{j+1}^{(1)}$ a partir de algún método multipasos explícito. Cuando un sistema de ecuaciones rígidas participa, se requiere una generalización ya sea para el método de Newton o de la secante. Estos temas se consideran en el capítulo 10.

ALGORITMO 5.8

Iteración trapezoidal con Newton

Para aproximar la solución del problema de valor inicial

$$y' = f(t, y), \quad \text{para } a \leq t \leq b, \text{ con } y(a) = \alpha$$

en $(N+1)$ números igualmente espaciados en el intervalo $[a, b]$:

ENTRADA extremos a, b; entero N; condición inicial α; tolerancia TOL; número máximo de iteraciones M en cualquier paso uno.

SALIDA aproximación w para y en los valores $(N+1)$ de t o un mensaje de falla.

Paso 1 Determine $h = (b-a)/N$;
$t = a$;
$w = \alpha$;
SALIDA (t, w).

Paso 2 Para $i = 1, 2, \ldots, N$ haga los pasos 3–7.

Paso 3 Determine $k_1 = w + \frac{h}{2} f(t, w)$;
$w_0 = k_1$;
$j = 1$;
$FLAG = 0$.

Paso 4 Mientras $FLAG = 0$ haga los pasos 5–6.

Paso 5 Determine $w = w_0 - \dfrac{w_0 - \frac{h}{2} f(t+h, w_0) - k_1}{1 - \frac{h}{2} f_y(t+h, w_0)}$.

Paso 6 Si $|w - w_0| < TOL$ entonces determine $FLAG = 1$
entonces determine $j = j + 1$;
$w_0 = w$;
si $j > M$ entonces
SALIDA ('El número máximo de iteraciones excedido');
PARE.

Paso 7 Determine $t = a + ih$;
SALIDA (t, w). *Fin del paso 2*

Paso 8 PARE. ■

Ilustración El problema de valor inicial rígido

$$y' = 5e^{5t}(y - t)^2 + 1, \quad 0 \le t \le 1, \quad y(0) = -1$$

tiene como solución $y(t) = t - e^{-5t}$. Para mostrar los efectos de rigidez, el método trapezoidal implícito y el método Runge-Kutta de cuarto orden se aplican con $N = 4$, al hacer $h = 0.25$, y con $N = 5$, que da $h = 0.20$.

El método trapezoidal se desempeña correctamente en ambos casos, usando $M = 10$ y $TOL = 10^{-6}$, al igual que Runge-Kutta con $h = 0.2$. Sin embargo, $h = 0.25$ está fuera de la región de estabilidad absoluta del método Runge-Kutta, que es evidente a partir de los resultados en la tabla 5.24. ■

Tabla 5.24

	Método Runge-Kutta		Método trapezoidal	
	$h = 0.2$		$h = 0.2$	
t_i	w_i	$\lvert y(t_i) - w_i \rvert$	w_i	$\lvert y(t_i) - w_i \rvert$
0.0	−1.0000000	0	−1.0000000	0
0.2	−0.1488521	1.9027×10^{-2}	−0.1414969	2.6383×10^{-2}
0.4	0.2684884	3.8237×10^{-3}	0.2748614	1.0197×10^{-2}
0.6	0.5519927	1.7798×10^{-3}	0.5539828	3.7700×10^{-3}
0.8	0.7822857	6.0131×10^{-4}	0.7830720	1.3876×10^{-3}
1.0	0.9934905	2.2845×10^{-4}	0.9937726	5.1050×10^{-4}
	$h = 0.25$		$h = 0.25$	
t_i	w_i	$\lvert y(t_i) - w_i \rvert$	w_i	$\lvert y(t_i) - w_i \rvert$
0.0	−1.0000000	0	−1.0000000	0
0.25	0.4014315	4.37936×10^{-1}	0.0054557	4.1961×10^{-2}
0.5	3.4374753	3.01956×10^{0}	0.4267572	8.8422×10^{-3}
0.75	1.44639×10^{23}	1.44639×10^{23}	0.7291528	2.6706×10^{-3}
1.0	Desbordado		0.9940199	7.5790×10^{-4}

Aquí hemos presentado sólo una breve introducción para lo que el lector, que con frecuencia encuentra ecuaciones diferenciales rígidas, debería saber. Para mayores detalles consulte [Ge2], [Lam] o [SGe].

La sección Conjunto de ejercicios 5.11 está disponible en línea. Encuentre la ruta de acceso en las páginas preliminares.

5.12 Software numérico

La biblioteca IMSL incluye dos subrutinas para aproximar las soluciones de los problemas de valor inicial. Cada uno de los métodos resuelve un sistema de *m* ecuaciones de primer orden en *m* variables. Las ecuaciones son de la forma

$$\frac{du_i}{dt} = f_i(t, u_1, u_2, \ldots, u_m), \quad \text{para } i = 1, 2, \ldots, m,$$

donde $u_i(t_0)$ se da para cada i. Una subrutina de tamaño de paso variable está basada en los métodos Runge-Kutta de cuarto y quinto orden descritos en el ejercicio 7 de la sección 5.5. Una subrutina de tipo Adams también está disponible para utilizarse para ecuaciones rígidas con base en un método de C. William Gear. Este método utiliza métodos multipasos implícitos de orden hasta 12 y fórmulas de diferenciación regresiva de orden hasta 5.

Los procedimientos tipo Runge-Kutta contenidos en la biblioteca NAG están basados en la forma Merson del método Runge-Kutta. Un método Adams de una variable y de tamaño de paso variable para sistemas rígidos. Otras rutinas incluyen los mismos métodos, pero se iteran hasta que un componente de la solución logra un valor determinado o hasta que una función de la solución es cero.

La biblioteca netlib incluye varias subrutinas para aproximar las soluciones de los problemas de valor inicial en el paquete EDO. Una subrutina está basada en los métodos Runge-Kutta-Verner de quinto y sexto orden, otra en los métodos Runge-Kutta-Fehlberg de cuarto y quinto orden, como se describe en la página 220 de la sección 5.5. Una subrutina para problemas de valor inicial de ecuación diferencial ordinaria rígida está basada en una fórmula de diferenciación regresiva de coeficiente variable.

Las secciones Preguntas de análisis, Conceptos clave y Revisión del capítulo están disponibles en línea. Encuentre la ruta de acceso en las páginas preliminares.

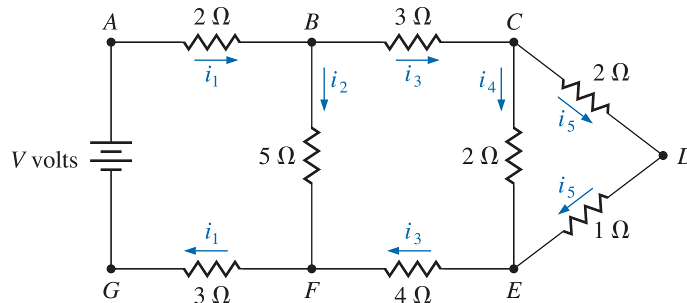
CAPÍTULO

6 Métodos directos para resolver sistemas lineales

Introducción

Las leyes de Kirchhoff de circuitos eléctricos establecen que tanto el flujo de corriente de la red hacia cada unión como la caída de voltaje alrededor de cada ciclo cerrado de un circuito son cero. Suponga que se aplica un potencial de V volts entre los puntos A y G en el circuito y que i_1, i_2, i_3, i_4 y i_5 representan el flujo de corriente como se muestra en el diagrama. Usando G como punto de referencia, las leyes de Kirchhoff implican que las corrientes satisfacen el siguiente sistema de ecuaciones lineales:

$$\begin{aligned} 5i_1 + 5i_2 &= V, \\ i_3 - i_4 - i_5 &= 0, \\ 2i_4 - 3i_5 &= 0, \\ i_1 - i_2 - i_3 &= 0, \\ 5i_2 - 7i_3 - 2i_4 &= 0. \end{aligned}$$



En este capítulo se considerará la solución de sistemas de este tipo. Esta aplicación se analiza en el ejercicio 23 de la sección 6.6.

Los sistemas de ecuaciones lineales están relacionados con muchos problemas en ingeniería y ciencia, así como con aplicaciones de matemáticas para las ciencias sociales y el estudio cuantitativo de problemas de negocios y economía.

En este capítulo, consideramos *métodos directos* para resolver un sistema de n ecuaciones lineales en n variables. Este sistema tiene la forma

$$\begin{aligned} E_1: &\quad a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n = b_1, \\ E_2: &\quad a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n = b_2, \\ &\qquad \vdots \\ E_n: &\quad a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n = b_n. \end{aligned} \tag{6.1}$$

En este sistema nos proporcionan las constantes a_{ij}, para cada $i, j = 1, 2, \ldots, n$, y b_i, para cada $i = 1, 2, \ldots, n$ y necesitamos determinar las incógnitas $x_1, \ldots, x_n$.

Las técnicas directas son métodos que proporcionan teóricamente la solución exacta del sistema en un número finito de pasos. En la práctica, por supuesto, la solución obtenida estará contaminada por el error de redondeo relacionado con la aritmética utilizada. Analizar el efecto de este error de redondeo y determinar formas para mantenerlo bajo control será un componente muy importante de este capítulo.

No se supone que un curso de álgebra lineal sea un prerrequisito para este capítulo, por lo que incluiremos una serie de nociones básicas sobre el tema. Estos resultados también se utilizarán en el capítulo 7, donde consideraremos los métodos para aproximar la solución de los sistemas lineales con métodos iterativos.

6.1 Sistemas de ecuaciones lineales

Utilizamos tres operaciones para simplificar el sistema lineal dado en la ecuación (6.1):

1. La ecuación E_i puede multiplicarse por cualquier constante λ diferente de cero y la ecuación resultante puede usarse en lugar de E_i. Esta operación se denota como $(\lambda E_i) \to (E_i)$.
2. La ecuación E_j puede multiplicarse por cualquier constante λ diferente de cero y sumarse con la ecuación E_i y la ecuación resultante puede usarse en lugar de E_i. Esta operación se denota como $(E_i + \lambda E_j) \to (E_i)$.
3. El orden de las ecuaciones E_i y E_j puede intercambiarse. Esta operación se denota $(E_i) \leftrightarrow (E_j)$.

Mediante una secuencia de estas operaciones, un sistema lineal se transformará de manera constante en uno nuevo más fácil de resolver y con las mismas soluciones (consulte el ejercicio 13). La secuencia de las operaciones se ilustra a continuación.

Ilustración Las cuatro ecuaciones

$$\begin{aligned}
E_1 &: \quad x_1 + x_2 \phantom{{}- x_3} + 3x_4 = 4, \\
E_2 &: \quad 2x_1 + x_2 - x_3 + x_4 = 1, \\
E_3 &: \quad 3x_1 - x_2 - x_3 + 2x_4 = -3, \\
E_4 &: \quad -x_1 + 2x_2 + 3x_3 - x_4 = 4,
\end{aligned} \tag{6.2}$$

se resolverán para x_1, x_2, x_3 y x_4. Primero usaremos la ecuación E_1 para eliminar el valor desconocido x_1 de las ecuaciones E_2, E_3 y E_4) al realizar $(E_2 - 2E_1) \to (E_2)$, $(E_3 - 3E_1) \to (E_3)$, y $(E_4 + E_1) \to (E_4)$. Por ejemplo, en la segunda ecuación

$$(E_2 - 2E_1) \to (E_2)$$

produce

$$(2x_1 + x_2 - x_3 + x_4) - 2(x_1 + x_2 + 3x_4) = 1 - 2(4),$$

que se simplifica para el resultado mostrado como E_2 en

$$\begin{aligned}
E_1 &: \quad x_1 + x_2 \phantom{{}- x_3} + 3x_4 = 4, \\
E_2 &: \quad - x_2 - x_3 - 5x_4 = -7, \\
E_3 &: \quad - 4x_2 - x_3 - 7x_4 = -15, \\
E_4 &: \quad 3x_2 + 3x_3 + 2x_4 = 8.
\end{aligned}$$

Por simplicidad, las ecuaciones nuevas se etiquetan otra vez como E_1, E_2, E_3 y E_4.

En el nuevo sistema, E_2 se usa para eliminar la incógnita x_2 de E_3 y E_4 al realizar $(E_3 - 4E_2) \to (E_3)$ y $(E_4 + 3E_2) \to (E_4)$. Esto resulta en

$$\begin{aligned} E_1 : &\quad x_1 + x_2 \qquad\quad + \ 3x_4 = \ \ 4, \\ E_2 : &\quad \quad - x_2 - \ x_3 - \ 5x_4 = -7, \\ E_3 : &\quad \qquad\qquad 3x_3 + 13x_4 = \ \ 13, \\ E_4 : &\quad \qquad\qquad\quad - 13x_4 = -13. \end{aligned} \tag{6.3}$$

Ahora, el sistema de ecuaciones (6.3) tiene una **forma triangular** (o **reducida**) y se puede resolver para las incógnitas mediante un **proceso de sustitución hacia atrás**. Ya que E_4 implica $x_4 = 1$, podemos resolver E_3 para x_3 para obtener

$$x_3 = \frac{1}{3}(13 - 13x_4) = \frac{1}{3}(13 - 13) = 0.$$

Al continuar, E_2 nos da

$$x_2 = -(-7 + 5x_4 + x_3) = -(-7 + 5 + 0) = 2,$$

y E_1 da

$$x_1 = 4 - 3x_4 - x_2 = 4 - 3 - 2 = -1.$$

La solución para el sistema (6.3) y, por consiguiente, para el sistema (6.2) es, por lo tanto $x_1 = -1$, $x_2 = 2$, $x_3 = 0$, y $x_4 = 1$. ■

Matrices y vectores

Al realizar los cálculos en la ilustración, no es necesario escribir las ecuaciones completas en cada paso o retener las variables x_1, x_2, x_3, y x_4 a lo largo de los cálculos, si siempre permanecen en la misma columna. La única variación de sistema a sistema se presenta en los coeficientes de las incógnitas y en los valores del lado derecho de las ecuaciones. Debido a esto, a menudo, se reemplaza un sistema lineal con una *matriz* que contiene toda la información sobre el sistema necesaria para determinar su solución, pero de manera compacta y una que se representa fácilmente en una computadora.

Definición 6.1 Una **matriz $n \times m$ (n por m)** es un arreglo rectangular de elementos con n filas y m columnas en las que no sólo se encuentra el valor de un elemento importante, sino también su posición en el arreglo. ■

La notación para una matriz $n \times m$ será una mayúscula, como A, para la matriz y minúsculas con subíndices dobles, como a_{ij}, para referirse a la entrada en la intersección de la i-ésima fila y la j-ésima columna; es decir

$$A = [a_{ij}] = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nm} \end{bmatrix}.$$

Ejemplo 1 Determine el tamaño y las entradas respectivas de la matriz.

$$A = \begin{bmatrix} 2 & -1 & 7 \\ 3 & 1 & 0 \end{bmatrix}.$$

Solución La matriz tiene dos filas y tres columnas, por lo que su tamaño es 2×3. Sus entradas se describen con $a_{11} = 2$, $a_{12} = -1$, $a_{13} = 7$, $a_{21} = 3$, $a_{22} = 1$, y $a_{23} = 0$. ■

La matriz 1 = n

$$A = [a_{11} \;\; a_{12} \;\; \cdots \;\; a_{1n}]$$

recibe el nombre de **vector fila n-dimensional** y una matriz $n \times 1$

$$A = \begin{bmatrix} a_{11} \\ a_{21} \\ \vdots \\ a_{n1} \end{bmatrix}$$

recibe el nombre de **vector columna n-dimensional**. Normalmente, para los vectores se omiten los subíndices y se utilizan letras minúsculas negritas para denotarlos. Por lo tanto,

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

denota un vector de columna y

$$\mathbf{y} = [y_1 \;\; y_2 \;\; \ldots \;\; y_n]$$

un vector de fila. Además, a menudo, los vectores fila tienen comas entre las entradas para hacer que la separación sea clara. Por lo que usted podría ver $\boldsymbol{y}$ escrita como $\mathbf{y} = [y_1, \;\; y_2, \;\; \ldots, \;\; y_n]$.

Una matriz $n \times (n + 1)$ se puede usar para representar el sistema lineal

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= b_1, \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &= b_2, \\ \vdots \qquad\qquad & \quad \vdots \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n &= b_n, \end{aligned}$$

al construir primero

$$A = [a_{ij}] = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix} \quad \text{y} \quad \mathbf{b} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}$$

$$[A, \mathbf{b}] = \left[\begin{array}{cccc:c} a_{11} & a_{12} & \cdots & a_{1n} & b_1 \\ a_{21} & a_{22} & \cdots & a_{2n} & b_2 \\ \vdots & \vdots & & \vdots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} & b_n \end{array}\right],$$

Aumentada se refiere al hecho de que el lado derecho del sistema se ha incluido en la matriz.

donde la línea punteada vertical se usa para separar los valores de los coeficientes de las incógnitas con los del lado derecho de las ecuaciones. El arreglo $[A, \mathbf{b}]$ recibe el nombre de **matriz aumentada**.

La repetición de las operaciones implicadas en la ilustración de la página 270 con la notación resulta en considerar primero la matriz aumentada:

$$\left[\begin{array}{rrrr:r} 1 & 1 & 0 & 3 & 4 \\ 2 & 1 & -1 & 1 & 1 \\ 3 & -1 & -1 & 2 & -3 \\ -1 & 2 & 3 & -1 & 4 \end{array}\right].$$

Realizar las operaciones de acuerdo con lo descrito en el ejemplo produce matrices aumentadas

$$\left[\begin{array}{rrrr:r} 1 & 1 & 0 & 3 & 4 \\ 0 & -1 & -1 & -5 & -7 \\ 0 & -4 & -1 & -7 & -15 \\ 0 & 3 & 3 & 2 & 8 \end{array}\right] \quad \text{y} \quad \left[\begin{array}{rrrr:r} 1 & 1 & 0 & 3 & 4 \\ 0 & -1 & -1 & -5 & -7 \\ 0 & 0 & 3 & 13 & 13 \\ 0 & 0 & 0 & -13 & -13 \end{array}\right].$$

Una técnica similar a la eliminación gaussiana apareció por primera vez durante la dinastía Han en China, en el texto *Chapters on the Mathematical Art (Capítulos sobre el arte matemático),* escrito alrededor del año 200 a.C. Joseph Louis Lagrange (1736–1813) describió una técnica similar a este procedimiento en 1778 para el caso en que el valor de cada ecuación sea 0. Gauss proporcionó una descripción más general en *Theoria Motus corporum coelestium sectionibus solem ambientium*, que describía la técnica de mínimos cuadrados que usó en 1801 para determinar la órbita del planeta menor Ceres.

Ahora, la matriz final se puede transformar en su sistema lineal correspondiente y es posible obtener soluciones para x_1, x_2, x_3 y x_4. Este procedimiento recibe el nombre de **eliminación gaussiana con sustitución hacia atrás.**

El procedimiento general de eliminación gaussiana aplicado al sistema lineal

$$\begin{aligned} E_1: &\quad a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n = b_1, \\ E_2: &\quad a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n = b_2, \\ &\quad \vdots \qquad\qquad\qquad\qquad \vdots \\ E_n: &\quad a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n = b_n, \end{aligned} \tag{6.4}$$

se maneja de manera similar. Primero, forme la matriz aumentada $\tilde{A}$,

$$\tilde{A} = [A, \mathbf{b}] = \left[\begin{array}{cccc:c} a_{11} & a_{12} & \cdots & a_{1n} & a_{1,n+1} \\ a_{21} & a_{22} & \cdots & a_{2n} & a_{2,n+1} \\ \vdots & \vdots & & \vdots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} & a_{n,n+1} \end{array}\right], \tag{6.5}$$

donde A denota la matriz formada por los coeficientes. Las entradas en la $(n + 1)$ columna son los valores de $\boldsymbol{b}$; es decir, $a_{i,n+1} = b_i$ para cada $i = 1, 2, \ldots, n$.

Siempre y cuando $a_{11} \neq 0$, realizamos las operaciones correspondientes para

$$(E_j - (a_{j1}/a_{11})E_1) \to (E_j) \quad \text{para cada } j = 2, 3, \ldots, n$$

para eliminar el coeficiente de x_1 en cada una de estas filas. A pesar de que se espera que cambien las entradas en las filas $2, 3, \ldots, n$, para facilidad de notación, nuevamente denotamos la entrada en la i-ésima fila y la j-ésima columna mediante a_{ij}. Con esto en mente, seguimos el procedimiento secuencial para $i = 2, 3, \ldots, n - 1$ y realizamos la operación

$$(E_j - (a_{ji}/a_{ii})E_i) \to (E_j) \quad \text{para cada } j = i + 1, i + 2, \ldots, n,$$

siempre y cuando $a_{ii} \neq 0$. Esto elimina (cambia el coeficiente a cero) x_i en cada fila debajo del i-ésimo para todos los valores de $i = 1, 2, \ldots, n - 1$. La matriz resultante tiene la forma

$$\tilde{\tilde{A}} = \left[\begin{array}{cccc:c} a_{11} & a_{12} & \cdots & a_{1n} & a_{1,n+1} \\ 0 & a_{22} & \cdots & a_{2n} & a_{2,n+1} \\ \vdots & \ddots & \ddots & \vdots & \vdots \\ 0 & \cdots & 0 & a_{nn} & a_{n,n+1} \end{array}\right],$$

donde, excepto en la primera columna, no se espera que los valores de a_{ij} concuerden con los de la matriz original $\tilde{A}$. La matriz $\tilde{\tilde{A}}$ representa un sistema lineal con la misma solución establecida como el sistema original.

El sistema lineal nuevo es triangular,

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= a_{1,n+1}, \\ a_{22}x_2 + \cdots + a_{2n}x_n &= a_{2,n+1}, \\ \vdots \qquad\quad & \quad \vdots \\ a_{nn}x_n &= a_{n,n+1}, \end{aligned}$$

Ejemplo 2 Represente el sistema lineal

$$\begin{aligned} E_1: &\quad x_1 - x_2 + 2x_3 - x_4 = -8,\\ E_2: &\quad 2x_1 - 2x_2 + 3x_3 - 3x_4 = -20,\\ E_3: &\quad x_1 + x_2 + x_3 = -2,\\ E_4: &\quad x_1 - x_2 + 4x_3 + 3x_4 = 4, \end{aligned}$$

como una matriz aumentada y utilice la eliminación gaussiana para encontrar su solución.

Solución La matriz aumentada es

$$\tilde{A} = \tilde{A}^{(1)} = \left[\begin{array}{rrrr:r} 1 & -1 & 2 & -1 & -8\\ 2 & -2 & 3 & -3 & -20\\ 1 & 1 & 1 & 0 & -2\\ 1 & -1 & 4 & 3 & 4 \end{array}\right].$$

Al realizar las operaciones

$$(E_2 - 2E_1) \to (E_2),\ (E_3 - E_1) \to (E_3), \quad \text{y} \quad (E_4 - E_1) \to (E_4),$$

obtenemos

$$\tilde{A}^{(2)} = \left[\begin{array}{rrrr:r} 1 & -1 & 2 & -1 & -8\\ 0 & 0 & -1 & -1 & -4\\ 0 & 2 & -1 & 1 & 6\\ 0 & 0 & 2 & 4 & 12 \end{array}\right].$$

El elemento pivote para una columna específica es la entrada utilizada para colocar ceros en las otras entradas en esa columna.

La entrada diagonal $a_{22}^{(2)}$, conocida como **elemento pivote**, es 0, por lo que el procedimiento no puede continuar en su forma actual. Sin embargo se permiten las operaciones $(E_i) \leftrightarrow (E_j)$, por lo que se realiza una búsqueda de elementos $a_{32}^{(2)}$ y $a_{42}^{(2)}$ para el primer elemento diferente de cero. Puesto que $a_{32}^{(2)} \neq 0$, se realiza la operación $(E_2) \leftrightarrow (E_3)$ para obtener una matriz nueva

$$\tilde{A}^{(2)'} = \left[\begin{array}{rrrr:r} 1 & -1 & 2 & -1 & -8\\ 0 & 2 & -1 & 1 & 6\\ 0 & 0 & -1 & -1 & -4\\ 0 & 0 & 2 & 4 & 12 \end{array}\right].$$

Puesto que x_2 ya se ha eliminado de E_3 y $\tilde{A}^{(3)}$ será $\tilde{A}^{(2)'}$, y los cálculos continúan con la operación $(E_4 + 2E_3) \to (E_4)$, lo que nos da

$$\tilde{A}^{(4)} = \left[\begin{array}{rrrr:r} 1 & -1 & 2 & -1 & -8\\ 0 & 2 & -1 & 1 & 6\\ 0 & 0 & -1 & -1 & -4\\ 0 & 0 & 0 & 2 & 4 \end{array}\right].$$

Finalmente, la matriz se vuelve a convertir en un sistema lineal que tiene una solución equivalente a la solución del sistema original y se aplica la sustitución hacia atrás:

$$\begin{aligned} x_4 &= \frac{4}{2} = 2,\\ x_3 &= \frac{[-4 - (-1)x_4]}{-1} = 2,\\ x_2 &= \frac{[6 - [(-1)x_3 + x_4]]}{2} = 3,\\ x_1 &= \frac{[-8 - [(-1)x_2 + 2x_3 + (-1)x_4]]}{1} = -7. \end{aligned}$$

■

El ejemplo 2 ilustra lo que se hace si $a_{kk}^{(k)} = 0$ para algunas $k = 1, 2, \ldots, n-1$. Para la k-ésima columna de $\tilde{A}^{(k-1)}$ desde la k-ésima fila hasta la enésima fila se busca la primera entrada diferente a cero. Si $a_{pk}^{(k)} \neq 0$ para algunas p, con $k+1 \leq p \leq n$, entonces se realiza la operación $(E_k) \leftrightarrow (E_p)$ para obtener $\tilde{A}^{(k-1)'}$. El procedimiento puede continuar para formar $\tilde{A}^{(k)}$ y así sucesivamente. Si $a_{pk}^{(k)} = 0$ para cada p, se puede mostrar (consulte el teorema 6.17 en la página 298) que el sistema lineal no tiene una solución única y el procedimiento se detiene. Finalmente, si $a_{nn}^{(n)} = 0$, el sistema lineal no tiene una solución única y, de nuevo, el procedimiento se detiene.

El algoritmo 6.1 resume la eliminación gaussiana con sustitución hacia atrás. El algoritmo incorpora el pivote cuando uno de los pivotes $a_{kk}^{(k)}$ es 0 al intercambiar la k-ésima fila con la p-ésima fila, donde p es el entero más pequeño superior a k para el que $a_{pk}^{(k)} \neq 0$.

ALGORITMO
6.1

Eliminación gaussiana con sustitución hacia atrás

Para resolver el sistema lineal $n \times n$

$$
\begin{aligned}
E_1: \quad & a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n = a_{1,n+1} \\
E_2: \quad & a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n = a_{2,n+1} \\
\vdots \quad & \quad \vdots \qquad\quad \vdots \qquad\qquad\quad \vdots \qquad\quad \vdots \\
E_n: \quad & a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n = a_{n,n+1}
\end{aligned}
$$

ENTRADA número de incógnitas y ecuaciones n; matriz aumentada $A = [a_{ij}]$, donde $1 \leq i \leq n$ y $1 \leq j \leq n+1$.

SALIDA Para $x_1, x_2, \ldots, x_n$ o mensaje de que el sistema lineal no tiene solución única.

Paso 1 Para $i = 1, \ldots, n-1$ haga los pasos 2–4. (*Proceso de eliminación*.)

Paso 2 Si p es el entero más pequeño con $i \leq p \leq n$ y $a_{pi} \neq 0$.
si no es posible encontrar un entero p
entonces SALIDA ('no existe una solución única');
PARE.

Paso 3 Si $p \neq i$ entonces realice $(E_p) \leftrightarrow (E_i)$.

Paso 4 Para $j = i+1, \ldots, n$ haga los pasos 5 y 6.

Paso 5 Determine $m_{ji} = a_{ji}/a_{ii}$.

Paso 6 Ejecute $(E_j - m_{ji}E_i) \rightarrow (E_j)$;

Paso 7 Si $a_{nn} = 0$ entonces SALIDA ('no existe una solución única');
PARE.

Paso 8 Determine $x_n = a_{n,n+1}/a_{nn}$. (*Inicia sustitución hacia atrás*.)

Paso 9 Para $i = n-1, \ldots, 1$ determine $x_i = \left[a_{i,n+1} - \sum_{j=i+1}^{n} a_{ij}x_j\right] \Big/ a_{ii}$.

Paso 10 SALIDA $(x_1, \ldots, x_n)$; (*Procedimiento completado con éxito*.)
PARE. ■

Ilustración El objetivo de esta ilustración es mostrar lo que pasa si el algoritmo 6.1 falla. Los cálculos se realizarán de manera simultánea en dos sistemas lineales

$$\begin{aligned} x_1 + x_2 + x_3 &= 4, \\ 2x_1 + 2x_2 + x_3 &= 6, \\ x_1 + x_2 + 2x_3 &= 6, \end{aligned} \qquad \text{y} \qquad \begin{aligned} x_1 + x_2 + x_3 &= 4, \\ 2x_1 + 2x_2 + x_3 &= 4, \\ x_1 + x_2 + 2x_3 &= 6. \end{aligned}$$

Estos sistemas producen las matrices aumentadas

$$\tilde{A} = \left[\begin{array}{ccc:c} 1 & 1 & 1 & 4 \\ 2 & 2 & 1 & 6 \\ 1 & 1 & 2 & 6 \end{array}\right] \qquad \text{y} \qquad \tilde{A} = \left[\begin{array}{ccc:c} 1 & 1 & 1 & 4 \\ 2 & 2 & 1 & 4 \\ 1 & 1 & 2 & 6 \end{array}\right].$$

Puesto que $a_{11} = 1$, realizamos $(E_2 - 2E_1) \to (E_2)$ y $(E_3 - E_1) \to (E_3)$ para producir

$$\tilde{A} = \left[\begin{array}{ccc:c} 1 & 1 & 1 & 4 \\ 0 & 0 & -1 & -2 \\ 0 & 0 & 1 & 2 \end{array}\right] \qquad \text{y} \qquad \tilde{A} = \left[\begin{array}{ccc:c} 1 & 1 & 1 & 4 \\ 0 & 0 & -1 & -4 \\ 0 & 0 & 1 & 2 \end{array}\right].$$

En este punto, $a_{22} = a_{32} = 0$. El algoritmo requiere que el procedimiento se detenga y no se obtiene ninguna solución para el sistema. Al escribir las ecuaciones para cada sistema obtenemos

$$\begin{aligned} x_1 + x_2 + x_3 &= 4, \\ -x_3 &= -2, \\ x_3 &= 2, \end{aligned} \qquad \text{y} \qquad \begin{aligned} x_1 + x_2 + x_3 &= 4, \\ -x_3 &= -4, \\ x_3 &= 2. \end{aligned}$$

El primer sistema lineal tiene un número infinito de soluciones, que se pueden describir mediante $x_3 = 2$, $x_2 = 2 - x_1$, y x_1 de manera arbitraria.

El segundo sistema conduce a la contradicción $x_3 = 2$ y $x_3 = 4$, por lo que no existe solución. Sin embargo, en este caso, no hay solución *única*, como concluimos a partir del algoritmo 6.1. ■

A pesar de que el algoritmo 6.1 puede ser observado como las construcciones de las matrices aumentadas $\tilde{A}^{(1)}, \ldots, \tilde{A}^{(n)}$, los cálculos pueden realizarse con sólo un arreglo $n \times (n+1)$. En cada paso, simplemente reemplazamos el valor anterior de a_{ij} por el nuevo. Además, podemos almacenar multiplicadores de m_{ji} en las ubicaciones de a_{ji} porque a_{ji} tiene el valor 0 para cada $i = 1, 2, \ldots, n-1$ y $j = i+1, i+2, \ldots, n$. Por lo tanto, A se puede sobreescribir mediante los multiplicadores en las entradas debajo de la diagonal principal (es decir, las entradas de la forma a_{ji}, con $j > i$) y mediante las entradas recientemente calculadas de $\tilde{A}^{(n)}$ en y sobre la diagonal principal (las entradas de la forma a_{ij} con $j \leq i$). Estos valores se pueden obtener para resolver otros sistemas lineales relacionados con la matriz original A, como observaremos en la sección 6.5.

Conteo de operaciones

Tanto el tiempo requerido para completar los cálculos como el error de redondeo subsecuente dependen del número de operaciones aritméticas de punto flotante necesarias para resolver un problema de rutina. En general, la cantidad de tiempo requerido para realizar una multiplicación o división en una computadora es aproximadamente el mismo y es muy superior al requerido para realizar una suma o una resta. Las diferencias reales en el tiempo de ejecución, sin embargo, dependen del sistema computacional particular. Para demostrar las operaciones de conteo para un método determinado, contaremos las operaciones requeridas para resolver un sistema lineal común de n ecuaciones con n incógnitas mediante el algoritmo 6.1. Mantendremos el conteo de las sumas/restas separado del conteo de las multiplicaciones/divisiones debido al diferencial de tiempo.

No se realizan operaciones aritméticas hasta los pasos 5 y 6 en el algoritmo. El paso 5 requiere realizar $(n - i)$ divisiones. El reemplazo de la ecuación E_j mediante $(E_j - m_{ji} E_i)$ en el paso 6 requiere multiplicar m_{ji} por cada término en E_i, lo cual resulta en un total de $(n-i)(n-i+1)$ multiplicaciones. Después de completar esto, cada término de la ecuación resultante se resta del término correspondiente en E_j. Esto requiere $(n-i)(n-i+1)$ restas. Para cada $i = 1, 2, \ldots, n-1$, las operaciones requeridas en los pasos 5 y 6 son los siguientes.

Multiplicaciones/divisiones:

$$(n-i) + (n-i)(n-i+1) = (n-i)(n-i+2).$$

Sumas/restas:

$$(n-i)(n-i+1).$$

El número total de operaciones requeridas por los pasos 5 y 6 se obtiene al sumar el conteo de operaciones para cada i. Al recordar, a partir del cálculo que

$$\sum_{j=1}^{m} 1 = m, \quad \sum_{j=1}^{m} j = \frac{m(m+1)}{2}, \quad \text{y} \quad \sum_{j=1}^{m} j^2 = \frac{m(m+1)(2m+1)}{6},$$

tenemos el siguiente conteo de operaciones.

Multiplicaciones/divisiones:

$$\begin{aligned}
\sum_{i=1}^{n-1}(n-i)(n-i+2) &= \sum_{i=1}^{n-1}(n^2 - 2ni + i^2 + 2n - 2i) \\
&= \sum_{i=1}^{n-1}(n-i)^2 + 2\sum_{i=1}^{n-1}(n-i) = \sum_{i=1}^{n-1} i^2 + 2\sum_{i=1}^{n-1} i \\
&= \frac{(n-1)n(2n-1)}{6} + 2\frac{(n-1)n}{2} = \frac{2n^3 + 3n^2 - 5n}{6}.
\end{aligned}$$

Sumas/restas:

$$\begin{aligned}
\sum_{i=1}^{n-1}(n-i)(n-i+1) &= \sum_{i=1}^{n-1}(n^2 - 2ni + i^2 + n - i) \\
&= \sum_{i=1}^{n-1}(n-i)^2 + \sum_{i=1}^{n-1}(n-i) = \sum_{i=1}^{n-1} i^2 + \sum_{i=1}^{n-1} i \\
&= \frac{(n-1)n(2n-1)}{6} + \frac{(n-1)n}{2} = \frac{n^3 - n}{3}.
\end{aligned}$$

Los únicos otros pasos en el algoritmo 6.1 que implican operaciones aritméticas son los requeridos para la sustitución hacia atrás, los pasos 8 y 9 requieren una división. El paso 9 requiere $(n - i)$ multiplicaciones y $(n - i - 1)$ sumas para cada término de suma y, después, una resta y una división. El número total de operaciones en los pasos 8 y 9 es la siguiente.

Multiplicaciones/divisiones:

$$1 + \sum_{i=1}^{n-1}((n-i)+1) = 1 + \left(\sum_{i=1}^{n-1}(n-i)\right) + n - 1$$
$$= n + \sum_{i=1}^{n-1}(n-i) = n + \sum_{i=1}^{n-1} i = \frac{n^2+n}{2}.$$

Sumas/restas:

$$\sum_{i=1}^{n-1}((n-i-1)+1) = \sum_{i=1}^{n-1}(n-i) = \sum_{i=1}^{n-1} i = \frac{n^2-n}{2}$$

El número total de operaciones aritméticas en el algoritmo 6.1 es, por lo tanto:

Multiplicaciones/divisiones:

$$\frac{2n^3+3n^2-5n}{6} + \frac{n^2+n}{2} = \frac{n^3}{3} + n^2 - \frac{n}{3}.$$

Sumas/restas:

$$\frac{n^3-n}{3} + \frac{n^2-n}{2} = \frac{n^3}{3} + \frac{n^2}{2} - \frac{5n}{6}.$$

Para n grande, el número total de multiplicaciones y divisiones es aproximadamente $n^3/3$, ya que es el número total de sumas y restas. Por lo tanto, la cantidad de cálculos y el tiempo requerido aumenta con n en proporción a n^3, como se muestra en la tabla 6.1.

Tabla 6.1

n	Multiplicaciones/divisiones	Sumas/restas
3	17	11
10	430	375
50	44 150	42 875
100	343 300	338 250

La sección Conjunto de ejercicios 6.1 está disponible en línea. Encuentre la ruta de acceso en las páginas preliminares.

6.2 Estrategias de pivoteo

Al derivar el algoritmo 6.1, encontramos que se necesitaba un intercambio de filas cuando uno de los elementos pivote $a_{kk}^{(k)}$ es cero. Este intercambio de filas tiene la forma $(E_k) \leftrightarrow (E_p)$, donde p es el entero más pequeño superior a k con $a_{pk}^{(k)} \neq 0$. Para reducir el error de redondeo, a menudo, es necesario realizar intercambios de filas incluso cuando los elementos pivote no son cero.

Si $a_{kk}^{(k)}$ es de magnitud pequeña, en comparación con $a_{jk}^{(k)}$, entonces la magnitud del multiplicador

$$m_{jk} = \frac{a_{jk}^{(k)}}{a_{kk}^{(k)}}$$

será mucho más grande que 1. El error de redondeo introducido en el cálculo de uno de los términos $a_{kl}^{(k)}$ se multiplica mediante m_{jk} al calcular $a_{jl}^{(k+1)}$, que compone el error original. Además, al realizar la sustitución hacia atrás para

$$x_k = \frac{a_{k,n+1}^{(k)} - \sum_{j=k+1}^{n} a_{kj}^{(k)}}{a_{kk}^{(k)}},$$

con un valor pequeño de $a_{kk}^{(k)}$, cualquier error en el numerador puede incrementar drásticamente debido a la división de $a_{kk}^{(k)}$. En nuestro ejemplo observaremos que incluso para sistemas pequeños, el error de redondeo puede dominar los cálculos.

Ejemplo 1 Aplique la eliminación gaussiana para el sistema

$$\begin{aligned} E_1: &\quad 0.003000x_1 + 59.14x_2 = 59.17 \\ E_2: &\quad 5.291x_1 - 6.130x_2 = 46.78, \end{aligned}$$

por medio de aritmética de cuatro dígitos con redondeo y compare los resultados con la solución exacta $x_1 = 10.00$ y $x_2 = 1.000$.

Solución El primer elemento pivote $a_{11}^{(1)} = 0.003000$, es pequeño y está relacionado con el multiplicador,

$$m_{21} = \frac{5.291}{0.003000} = 1763.6\overline{6},$$

se redondea al número más grande 1764. Al realizar $(E_2 - m_{21}E_1) \to (E_2)$ y el redondeo adecuado da el sistema

$$\begin{aligned} 0.003000x_1 + 59.14x_2 &\approx 59.17 \\ -104300x_2 &\approx -104400, \end{aligned}$$

en lugar del sistema exacto, que es

$$\begin{aligned} 0.003000x_1 + 59.14x_2 &= 59.17 \\ -104309.37\overline{6}x_2 &= -104309.37\overline{6}. \end{aligned}$$

La disparidad en las magnitudes de $m_{21}a_{13}$ y a_{23} ha introducido error de redondeo, pero el error de redondeo aún no se ha propagado. La sustitución hacia atrás produce

$$x_2 \approx 1.001,$$

que es una aproximación cercana al valor real $x_2 = 1.000$. Sin embargo, debido al pivote pequeño $a_{11} = 0.003000$,

$$x_1 \approx \frac{59.17 - (59.14)(1.001)}{0.003000} = -10.00$$

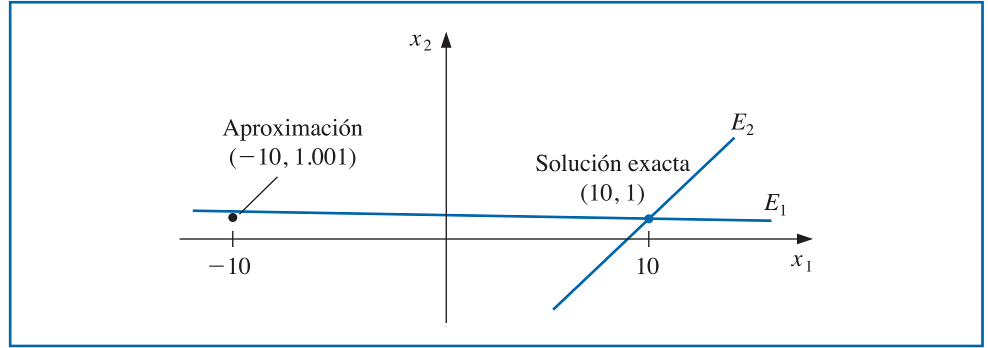
contiene el error pequeño de 0.001 multiplicado por

$$\frac{59.14}{0.003000} \approx 20000.$$

Esto arruina la aproximación del valor real $x_1 = 10.00$.

Esto es claramente un ejemplo artificial y la gráfica en la figura 6.1 muestra la razón por la que el error se puede presentar fácilmente. Para sistemas más grandes es mucho más difícil predecir cuándo se presentará un error de redondeo devastador. ■

Figura 6.1



Pivoteo parcial

El ejemplo 1 muestra la forma en la que pueden surgir dificultades cuando el elemento pivote $a_{kk}^{(k)}$ es pequeño relativo a las entradas $a_{ij}^{(k)}$, para $k \leq i \leq n$ y $k \leq j \leq n$. Para evitar este problema se realiza pivoteo al seleccionar un elemento $a_{pq}^{(k)}$ con una magnitud más grande como el pivote y al intercambiar las k-ésima y p-ésima filas. Esto puede ir seguido de un intercambio de la k-ésima y q-ésima columnas, en caso necesario.

La estrategia más simple, llamada pivoteo parcial, es seleccionar un elemento en la misma columna que está debajo de la diagonal y que tiene el máximo valor absoluto; específicamente, determinamos la $p \geq k$ más pequeña de tal forma que

$$|a_{pk}^{(k)}| = \max_{k \leq i \leq n} |a_{ik}^{(k)}|$$

y realizamos $(E_k) \leftrightarrow (E_p)$. En este caso, no se usa intercambio de columnas.

Ejemplo 2 Aplique la eliminación gaussiana para el sistema

$$\begin{aligned} E_1: &\quad 0.003000x_1 + 59.14x_2 = 59.17 \\ E_2: &\quad 5.291x_1 - 6.130x_2 = 46.78, \end{aligned}$$

con pivoteo parcial y aritmética de cuatro dígitos con redondeo y compare los resultados con la solución exacta $x_1 = 10.00$ y $x_2 = 1.000$.

Solución El procedimiento pivotal parcial requiere encontrar primero

$$\text{máx}\left\{|a_{11}^{(1)}|, |a_{21}^{(1)}|\right\} = \text{máx}\,\{|0.003000|, |5.291|\} = |5.291| = |a_{21}^{(1)}|.$$

Esto requiere realizar la operación $(E_2) \leftrightarrow (E_1)$ para producir el sistema equivalente

$$\begin{aligned} E_1: &\quad 5.291x_1 - 6.130x_2 = 46.78, \\ E_2: &\quad 0.003000x_1 + 59.14x_2 = 59.17. \end{aligned}$$

El multiplicador para este sistema es

$$m_{21} = \frac{a_{21}^{(1)}}{a_{11}^{(1)}} = 0.0005670,$$

y la operación $(E_2 - m_{21}E_1) \rightarrow (E_2)$ reduce el sistema a

$$\begin{aligned} 5.291x_1 - 6.130x_2 &\approx 46.78, \\ 59.14x_2 &\approx 59.14. \end{aligned}$$

La respuesta de cuatro dígitos resultante a partir de la sustitución hacia atrás son los valores correctos de $x_1 = 10.00$ y $x_2 = 1.000$. ■

La técnica recientemente descrita recibe el nombre de **pivoteo parcial** (o *pivoteo de columna máxima*) y se describe con detalle en el algoritmo 6.2. El intercambio de fila real se simula en el algoritmo al intercambiar los valores de *NROW* en el paso 5.

ALGORITMO
6.2

Eliminación gaussiana con pivoteo parcial

Para resolver el sistema lineal $n \times n$

$$
\begin{aligned}
E_1: &\quad a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n = a_{1,n+1}\\
E_2: &\quad a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n = a_{2,n+1}\\
&\quad \vdots \qquad\qquad\qquad\qquad\qquad \vdots\\
E_n: &\quad a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n = a_{n,n+1}
\end{aligned}
$$

ENTRADA número de incógnitas y ecuaciones n; matriz aumentada $A = [a_{ij}]$ donde $1 \leq i \leq n$ y $1 \leq j \leq n+1$.

SALIDA solución $x_1, \ldots, x_n$ o mensaje de que el sistema lineal no tiene solución única.

Paso 1 Para $i = 1, \ldots, n$ determine $NROW(i) = i$. (*Inicialice indicador de fila.*)

Paso 2 Para $i = 1, \ldots, n-1$ haga los pasos 3–6. (*Proceso de eliminación.*)

Paso 3 Sea p el entero más pequeño con $i \leq p \leq n$ y
$|a(NROW(p), i)| = \text{máx}_{i \leq j \leq n} |a(NROW(j), i)|$.
(*Notación:* $a(NROW(i), j) \equiv a_{NROW_i, j}$.)

Paso 4 Si $a(NROW(p), i) = 0$ entonces SALIDA ('no existe solución única');
PARE.

Paso 5 Si $NROW(i) \neq NROW(p)$ entonces determine $NCOPY = NROW(i)$;
$NROW(i) = NROW(p)$;
$NROW(p) = NCOPY$.
(*Intercambio de fila simulado.*)

Paso 6 Para $j = i+1, \ldots, n$ haga los pasos 7 y 8.

Paso 7 Determine $m(NROW(j), i) = a(NROW(j), i)/a(NROW(i), i)$.

Paso 8 Realice $(E_{NROW(j)} - m(NROW(j), i) \cdot E_{NROW(i)}) \rightarrow (E_{NROW(j)})$.

Paso 9 Si $a(NROW(n), n) = 0$ entonces SALIDA ('no existe solución única');
PARE.

Paso 10 Determine $x_n = a(NROW(n), n+1)/a(NROW(n), n)$.
(*Inicie sustitución hacia atrás.*)

Paso 11 Para $i = n-1, \ldots, 1$

determine $$x_i = \frac{a(NROW(i), n+1) - \sum_{j=i+1}^{n} a(NROW(i), j) \cdot x_j}{a(NROW(i), i)}.$$

Paso 12 SALIDA $(x_1, \ldots, x_n)$; (*Procedimiento completado con éxito.*)
PARE. ▪

Cada multiplicador m_{ji} en el algoritmo de pivoteo parcial tiene una magnitud menor o igual que 1. A pesar de que esta estrategia es suficiente para muchos sistemas lineales, surgen situaciones en las que es inadecuada.

Ilustración El sistema lineal

$$E_1: \quad 30.00x_1 + 591400x_2 = 591700,$$
$$E_2: \quad 5.291x_1 - \quad 6.130x_2 = 46.78,$$

es igual al de los ejemplos 1 y 2, excepto que todas las entradas en la primera ecuación se han multiplicado por 10^4. El procedimiento de pivoteo parcial descrito en el algoritmo 6.2 con aritmética de cuatro dígitos conduce a los mismos resultados obtenidos en el ejemplo 1. El valor máximo en la primera columna es 30.00 y el multiplicador

$$m_{21} = \frac{5.291}{30.00} = 0.1764$$

conduce al sistema

$$30.00x_1 + 591400x_2 \approx 591700,$$
$$-104300x_2 \approx -104400,$$

que tiene las mismas soluciones inadecuadas que el ejemplo 1: $x_2 \approx 1.001$ y $x_1 \approx -10.00$. ■

Pivoteo parcial escalado

El **pivoteo parcial escalado** (o *pivoteo de columna escalado*) es el adecuado para el sistema en la ilustración. Éste coloca al elemento en la posición pivote que es la más grande en relación con las entradas en su fila. El primer paso en este procedimiento es definir un factor de escala s_i para cada fila como

$$s_i = \max_{1\le j\le n} |a_{ij}|.$$

Si tenemos $s_i = 0$ para algunas i, entonces el sistema no tiene solución única ya que todas las entradas en la i-ésima fila son 0. Al suponer que éste no es el caso el intercambio de fila adecuado para colocar los ceros en la primera columna se determina al seleccionar el último entero p con

$$\frac{|a_{p1}|}{s_p} = \max_{1\le k\le n} \frac{|a_{k1}|}{s_k}$$

y realizar $(E_1) \leftrightarrow (E_p)$. El efecto de escalamiento es garantizar que el elemento más grande en cada columna tenga una magnitud *relativa* de 1 antes de la comparación realizada para determinar si se realiza el intercambio de fila.

De manera similar, antes de eliminar la variable x_i mediante las operaciones

$$E_k - m_{ki}E_i, \quad \text{para } k = i+1, \dots, n,$$

seleccionamos el entero más pequeño $p \ge i$ con

$$\frac{|a_{pi}|}{s_p} = \max_{i\le k\le n} \frac{|a_{ki}|}{s_k}$$

y realizamos el intercambio de fila $(E_i) \leftrightarrow (E_p)$ si $i \neq p$. Los factores de escala $s_1, \dots, s_n$ se calculan una sola vez, al inicio del procedimiento. Ahora, dependen de la fila, por lo que también se deben intercambiar cuando se realizan los intercambios de fila.

Ilustración Al aplicar el pivoteo parcial escalado para la ilustración previa obtenemos

$$s_1 = \text{máx}\{|30.00|, |591400|\} = 591400$$

y

$$s_2 = \text{máx}\{|5.291|, |-6.130|\} = 6.130.$$

Por consiguiente,

$$\frac{|a_{11}|}{s_1} = \frac{30.00}{591400} = 0.5073 \times 10^{-4}, \qquad \frac{|a_{21}|}{s_2} = \frac{5.291}{6.130} = 0.8631,$$

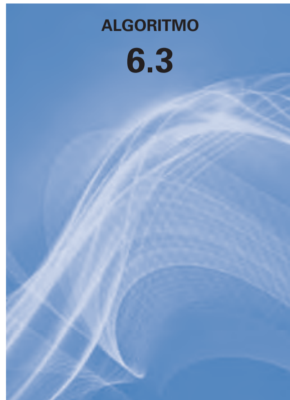
y se realiza el intercambio $(E_1) \leftrightarrow (E_2)$.

Al aplicar la eliminación gaussiana al nuevo sistema

$$5.291x_1 - 6.130x_2 = 46.78$$
$$30.00x_1 + 591400x_2 = 591700$$

produce los resultados correctos $x_1 = 10.00$ y $x_2 = 1.000$. ■

Algoritmo 6.3 implementa el pivoteo parcial escalado.



Eliminación gaussiana con pivoteo parcial escalado.

Los únicos pasos en este algoritmo que difieren de los del algoritmo 6.2 son:

Paso 1 Para $i = 1, \ldots, n$ determine $s_i = \text{máx}_{1\le j\le n} |a_{ij}|$;
si $s_i = 0$ entonces SALIDA ('no existe solución única');
PARE.
también determine $NROW(i) = i$.

Paso 2 Para $i = 1, \ldots, n-1$ haga los pasos 3–6. (*Proceso de eliminación.*)

Paso 3 Si p es el entero más pequeño con $i \le p \le n$ y

$$\frac{|a(NROW(p), i)|}{s(NROW(p))} = \max_{i\le j\le n} \frac{|a(NROW(j), i)|}{s(NROW(j))}.$$

■

Ejemplo 3 Resuelva el sistema lineal con aritmética de redondeo de tres dígitos.

$$\begin{aligned} 2.11x_1 - 4.21x_2 + 0.921x_3 &= 2.01,\\ 4.01x_1 + 10.2x_2 - 1.12x_3 &= -3.09,\\ 1.09x_1 + 0.987x_2 + 0.832x_3 &= 4.21.\end{aligned}$$

Solución Tenemos $s_1 = 4.21$, $s_2 = 10.2$, y $s_3 = 1.09$. Por lo que,

$$\frac{|a_{11}|}{s_1} = \frac{2.11}{4.21} = 0.501, \quad \frac{|a_{21}|}{s_1} = \frac{4.01}{10.2} = 0.393, \quad \text{y} \quad \frac{|a_{31}|}{s_3} = \frac{1.09}{1.09} = 1.$$

La matriz aumentada AA se define mediante

$$\left[\begin{array}{ccc:c} 2.11 & -4.21 & .921 & 2.01\\ 4.01 & 10.2 & -1.12 & -3.09\\ 1.09 & .987 & .832 & 4.21 \end{array}\right].$$

Puesto que $|a_{31}|/s_3$ es el más grande, realizamos $(E_1) \leftrightarrow (E_3)$ para obtener

$$\left[\begin{array}{ccc:c} 1.09 & .987 & .832 & 4.21 \\ 4.01 & 10.2 & -1.12 & -3.09 \\ 2.11 & -4.21 & .921 & 2.01 \end{array}\right].$$

Calcule los multiplicadores

$$m_{21} = \frac{a_{21}}{a_{11}} = 3.68; \quad m_{31} = \frac{a_{31}}{a_{11}} = 1.94.$$

Realice las primeras dos eliminaciones para producir

$$\left[\begin{array}{ccc:c} 1.09 & .987 & .832 & 4.21 \\ 0 & 6.57 & -4.18 & -18.6 \\ 0 & -6.12 & -.689 & -6.16 \end{array}\right].$$

Puesto que

$$\frac{|a_{22}|}{s_2} = \frac{6.57}{10.2} = 0.644 \quad \text{y} \quad \frac{|a_{32}|}{s_3} = \frac{6.12}{4.21} = 1.45,$$

realizamos $E_2 \leftrightarrow E_3$, lo cual nos da

$$\left[\begin{array}{ccc:c} 1.09 & .987 & .832 & 4.21 \\ 0 & -6.12 & -.689 & -6.16 \\ 0 & 6.57 & -4.18 & -18.6 \end{array}\right].$$

El multiplicador m_{32} se calcula a través de

$$m_{32} = \frac{a_{32}}{a_{22}} = -1.07,$$

y el siguiente paso de eliminación en la matriz

$$\left[\begin{array}{ccc:c} 1.09 & .987 & .832 & 4.21 \\ 0 & -6.12 & -.689 & -6.16 \\ 0 & 0 & -4.92 & -25.2 \end{array}\right].$$

Finalmente, la sustitución hacia atrás da la solución $\mathbf{x}$, la cual, para tres dígitos decimales, es $x_1 = -0.431$, $x_2 = 0.430$, y $x_3 = 5.12$. ■

Los primeros cálculos adicionales requeridos para pivoteo parcial escalado resultan de la determinación de los factores de escala; existen $(n-1)$ comparaciones para cada una de las n filas, para un total de

$$n(n-1) \text{ comparaciones.}$$

Para determinar el primer intercambio correcto, se realizan n divisiones, seguidas de $n-1$ comparaciones. Por lo que, la primera determinación de intercambio añade

$$n \text{ divisiones y } (n-1) \text{ comparaciones.}$$

Los factores de escalamiento se calculan solamente una vez, por lo que el segundo paso requiere

$$(n-1) \text{ divisiones y } (n-2) \text{ comparaciones.}$$

Seguimos de manera similar hasta obtener ceros por debajo de la diagonal principal en todas las filas, excepto enésima. El paso final requiere realizar

2 divisiones y 1 comparación.

En consecuencia, el pivoteo parcial escalado añade un total de

$$n(n-1)+\sum_{k=1}^{n-1} k = n(n-1)+\frac{(n-1)n}{2} = \frac{3}{2}n(n-1) \quad \text{comparaciones} \qquad (6.7)$$

y

$$\sum_{k=2}^{n} k = \left(\sum_{k=1}^{n} k\right) - 1 = \frac{n(n+1)}{2} - 1 = \frac{1}{2}(n-1)(n+2) \quad \text{divisiones}$$

al procedimiento de eliminación gaussiana. El tiempo requerido para realizar una comparación es aproximadamente el mismo que el de una suma/resta. Puesto que el tiempo total para realizar el procedimiento básico de eliminación gaussiana es $O(n^3/3)$ multiplicaciones/divisiones y $O(n^3/3)$ sumas/restas, el pivoteo parcial escalado no añade tiempo computacional significativo requerido para resolver un sistema para valores grandes de n.

Para enfatizar la importancia de seleccionar una sola vez los factores de escala, considere la cantidad de cálculos adicionales que se requerirían si se modificara el procedimiento de tal forma que se determinaran factores de escala nuevos cada vez que se realice una decisión de intercambio de fila. En este caso, el término $n(n-1)$ en la ecuación (6.7) sería reemplazado por

$$\sum_{k=2}^{n} k(k-1) = \frac{1}{3}n(n^2-1).$$

En consecuencia, esta técnica de pivoteo añadiría $O(n^3/3)$ comparaciones, además de las $[n(n+1)/2]-1$ divisiones.

Pivoteo completo

El pivoteo puede incluir el intercambio tanto de filas como de columnas. El **pivoteo completo** (o *máximo*) en el k-ésimo paso busca todas las entradas a_{ij}, para $i=k, k+1, \ldots, n$ y $j=k, k+1, \ldots, n$, para encontrar la entrada con mayor magnitud. Los intercambios de fila y de columna se realizan para colocar esta entrada en la posición pivote. El primer paso del pivoteo total requiere realizar n^2-1 comparaciones, el segundo paso requiere $(n-1)^2-1$ comparaciones, y así sucesivamente. El tiempo adicional total requerido para incorporar el pivoteo completo a la eliminación gaussiana es

$$\sum_{k=2}^{n}(k^2-1) = \frac{n(n-1)(2n+5)}{6}$$

comparaciones. El pivoteo completo es, por consiguiente, la estrategia recomendada solamente para sistemas en los que la precisión es fundamental y la cantidad de tiempo de ejecución necesario para este método se puede justificar.

La sección Conjunto de ejercicios 6.2 está disponible en línea. Encuentre la ruta de acceso en las páginas preliminares.

6.3 Álgebra lineal e inversión de matriz

Las matrices se introdujeron en la sección 6.1 como un método conveniente para expresar y manipular sistemas lineales. En esta sección consideramos cierta álgebra relacionada con matrices y muestra cómo se puede utilizar para resolver problemas asociados con sistemas lineales.

Definición 6.2 Dos matrices A y B son **iguales** si tienen el mismo número de filas y columnas, digamos, $n \times m$, y sí $a_{ij} = b_{ij}$, para cada $i = 1, 2, \ldots, n$ y $j = 1, 2, \ldots, m$. ■

Esta definición significa, por ejemplo, que

$$\begin{bmatrix} 2 & -1 & 7 \\ 3 & 1 & 0 \end{bmatrix} \neq \begin{bmatrix} 2 & 3 \\ -1 & 1 \\ 7 & 0 \end{bmatrix}$$

puesto que difieren en dimensión.

Aritmética de matriz

Dos operaciones importantes en matrices son la suma de dos matrices y la multiplicación de una matriz por un número real.

Definición 6.3 Si A y B son matrices, ambas de $n \times m$, la suma de A y B, denotada $A + B$, es la matriz $n \times m$ cuyas entradas son $a_{ij} + b_{ij}$, para cada $i = 1, 2, \ldots, n$ y $j = 1, 2, \ldots, m$. ■

Definición 6.4 Si A es una matriz $n \times m$ y λ es un número real, entonces la multiplicación escalar de λ y A, denotada λA, es la matriz $n \times m$ cuyas entradas son λa_{ij}, para cada $i = 1, 2, \ldots, n$ y $j = 1, 2, \ldots, m$. ■

Ejemplo 1 Determine $A + B$ y λA, cuando

$$A = \begin{bmatrix} 2 & -1 & 7 \\ 3 & 1 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 4 & 2 & -8 \\ 0 & 1 & 6 \end{bmatrix}, \quad \text{y } \lambda = -2.$$

Solución Tenemos

$$A + B = \begin{bmatrix} 2+4 & -1+2 & 7-8 \\ 3+0 & 1+1 & 0+6 \end{bmatrix} = \begin{bmatrix} 6 & 1 & -1 \\ 3 & 2 & 6 \end{bmatrix}$$

y

$$\lambda A = \begin{bmatrix} -2(2) & -2(-1) & -2(7) \\ -2(3) & -2(1) & -2(0) \end{bmatrix} = \begin{bmatrix} -4 & 2 & -14 \\ -6 & -2 & 0 \end{bmatrix}.$$ ■

Tenemos las siguientes propiedades generales para suma de matriz y multiplicación escalar. Estas propiedades son suficientes para clasificar el conjunto de todas las matrices $n \times m$ con entradas reales como **espacio vectorial** sobre el campo de números reales.

- Si dejamos que O denote una matriz, donde todas sus entradas son 0 y $-A$ denota la matriz cuyas entradas son $-a_{ij}$.

Teorema 6.5 Si A, B y C son matrices $n \times m$ y λ y μ son números reales. Se cumplen las siguientes propiedades de suma y multiplicación escalar:

i) $A + B = B + A$, **ii)** $(A + B) + C = A + (B + C)$,
iii) $A + O = O + A = A$, **iv)** $A + (-A) = -A + A = O$,
v) $\lambda(A + B) = \lambda A + \lambda B$, **vi)** $(\lambda + \mu)A = \lambda A + \mu A$,
vii) $\lambda(\mu A) = (\lambda\mu)A$, **viii)** $1A = A$.

Todas estas propiedades siguen resultados similares respecto a los números complejos. ■

Productos matriz-vector

El producto de matrices también se puede definir en ciertas instancias. Primero consideraremos el producto de una matriz $n \times m$ y un vector columna $m \times 1$.

Definición 6.6 Sea A una matriz $n \times m$ y $\mathbf{b}$ un vector columna m-dimensional. El **producto matriz-vector** de A y $\mathbf{b}$, denotado $A\mathbf{b}$, es un vector columna n-dimensional dado por

$$A\mathbf{b} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nm} \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{m} a_{1i} b_i \\ \sum_{i=1}^{m} a_{2i} b_i \\ \vdots \\ \sum_{i=1}^{m} a_{ni} b_i \end{bmatrix}.$$ ■

Para definir este producto, el número de columnas de la matriz A debe concordar con el número de filas del vector $\mathbf{b}$ y el resultado es otro vector columna con el número de filas que concuerda con el número de filas en la matriz.

Ejemplo 2 Determine el producto $A\mathbf{b}$ si $A = \begin{bmatrix} 3 & 2 \\ -1 & 1 \\ 6 & 4 \end{bmatrix}$ y $\mathbf{b} = \begin{bmatrix} 3 \\ -1 \end{bmatrix}$.

Solución Puesto que A tiene dimensión 3×2 y $\mathbf{b}$ tiene dimensión 2×1, el producto está definido y es un vector con tres filas. Éstas son

$$3(3) + 2(-1) = 7, \quad (-1)(3) + 1(-1) = -4, \quad \text{y} \quad 6(3) + 4(-1) = 14.$$

Es decir,

$$A\mathbf{b} = \begin{bmatrix} 3 & 2 \\ -1 & 1 \\ 6 & 4 \end{bmatrix} \begin{bmatrix} 3 \\ -1 \end{bmatrix} = \begin{bmatrix} 7 \\ -4 \\ 14 \end{bmatrix}.$$ ■

La introducción del producto matriz-vector nos permite ver el sistema lineal

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= b_1, \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &= b_2, \\ \vdots \qquad\qquad & \quad \vdots \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n &= b_n, \end{aligned}$$

como la ecuación matricial

$$A\mathbf{x} = \mathbf{b},$$

donde

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix}, \quad \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, \quad \text{y} \quad \mathbf{b} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix},$$

puesto que todas las entradas en el producto $A\mathbf{x}$ deben corresponder con las entradas en el vector **b.** Por lo tanto, la matriz $n \times m$ se puede ver como una función con dominio en conjunto de vectores columna m-dimensional y rango en un subconjunto de vectores columna n-dimensional.

Productos de matriz-matriz

Podemos utilizar la multiplicación matriz-vector para definir la multiplicación general matriz-matriz.

Definición 6.7 Si A es una matriz $n \times m$ y B una matriz $m \times p$. El **producto de la matriz** de A y B, denotado AB es una matriz C $n \times p$ cuyas entradas c_{ij} son

$$c_{ij} = \sum_{k=1}^{m} a_{ik}b_{kj} = a_{i1}b_{1j} + a_{i2}b_{2j} + \cdots + a_{im}b_{mj},$$

para cada $i = 1, 2, \ldots n$, y $j = 1, 2, \ldots, p$. ■

El cálculo de c_{ij} se puede observar como la multiplicación de las entradas de la i-ésima fila de A con entradas correspondientes en la j-ésima columna de B, seguida por la sumatoria; es decir,

$$[a_{i1}, a_{i2}, \ldots, a_{im}] \begin{bmatrix} b_{1j} \\ b_{2j} \\ \vdots \\ b_{mj} \end{bmatrix} = c_{ij},$$

donde

$$c_{ij} = a_{i1}b_{1j} + a_{i2}b_{2j} + \cdots + a_{im}b_{mj} = \sum_{k=1}^{m} a_{ik}b_{kj}.$$

Esto explica porqué el número de columnas de A debe ser igual al número de filas de B para el producto AB que se va a definir.

El siguiente ejemplo debería servir para aclarar el proceso de multiplicación de matrices.

Ejemplo 3 Determine todos los productos posibles de las matrices.

$$A = \begin{bmatrix} 3 & 2 \\ -1 & 1 \\ 1 & 4 \end{bmatrix}, \quad B = \begin{bmatrix} 2 & 1 & -1 \\ 3 & 1 & 2 \end{bmatrix},$$

$$C = \begin{bmatrix} 2 & 1 & 0 & 1 \\ -1 & 3 & 2 & 1 \\ 1 & 1 & 2 & 0 \end{bmatrix}, \quad \text{y} \quad D = \begin{bmatrix} 1 & -1 \\ 2 & -1 \end{bmatrix}.$$

Solución El tamaño de las matrices es

$$A : 3 \times 2, \quad B : 2 \times 3, \quad C : 3 \times 4, \quad \text{y} \quad D : 2 \times 2.$$

Los productos que se pueden definir y sus dimensiones son

$$AB : 3 \times 3, \quad BA : 2 \times 2, \quad AD : 3 \times 2, \quad BC : 2 \times 4, \quad DB : 2 \times 3, \quad \text{y} \quad DD : 2 \times 2.$$

Estos productos son

$$AB = \begin{bmatrix} 12 & 5 & 1 \\ 1 & 0 & 3 \\ 14 & 5 & 7 \end{bmatrix}, \qquad BA = \begin{bmatrix} 4 & 1 \\ 10 & 15 \end{bmatrix}, \qquad AD = \begin{bmatrix} 7 & -5 \\ 1 & 0 \\ 9 & -5 \end{bmatrix},$$

$$BC = \begin{bmatrix} 2 & 4 & 0 & 3 \\ 7 & 8 & 6 & 4 \end{bmatrix}, \qquad DB = \begin{bmatrix} -1 & 0 & -3 \\ 1 & 1 & -4 \end{bmatrix}, \qquad \text{y} \quad DD = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix}.$$

■

Observe que a pesar de que ambos productos de la matriz *AB* y *BA* están definidos, sus resultados son muy diferentes; ellos ni siquiera tienen la misma dimensión. En lenguaje matemático, decimos que la operación del producto de la matriz *no es conmutativo*; es decir, los productos en orden inverso pueden diferir. Éste es el caso incluso cuando ambos productos se definen y tienen la misma dimensión. Casi cualquier ejemplo mostrará esto, por ejemplo,

$$\begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 2 \\ 0 & 1 \end{bmatrix} \quad \text{mientras} \quad \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 2 & 1 \end{bmatrix}.$$

Ciertas operaciones importantes relacionadas con el producto de matrices se mantienen, como se indica en el siguiente resultado.

Teorema 6.8 Si A es una matriz $n \times m$, B es una matriz $m \times k$, C es una matriz $k \times p$, D es una matriz $m \times k$ y λ es un número real. Las siguientes propiedades se mantienen:

a) $A(BC) = (AB)C$; **b)** $A(B + D) = AB + AD$; **c)** $\lambda(AB) = (\lambda A)B = A(\lambda B)$.

Demostración La verificación de la propiedad en la parte a) se presenta para mostrar el método involucrado. Las otras partes se pueden demostrar de manera similar.

Para mostrar que $A(BC) = (AB)C$, calcule la ij-entrada de cada lado de la ecuación. BC es una matriz $m \times p$ con ij-entrada

$$(BC)_{sj} = \sum_{l=1}^{k} b_{sl} c_{lj}.$$

Por lo tanto, $A(BC)$ es una matriz $n \times p$ con entradas

$$[A(BC)]_{ij} = \sum_{s=1}^{m} a_{is} (BC)_{sj} = \sum_{s=1}^{m} a_{is} \left(\sum_{l=1}^{k} b_{sl} c_{lj} \right) = \sum_{s=1}^{m} \sum_{l=1}^{k} a_{is} b_{sl} c_{lj}.$$

De igual forma, AB es una matriz $n \times k$ con entradas

$$(AB)_{il} = \sum_{s=1}^{m} a_{is} b_{sl},$$

por lo que $(AB)C$ es una matriz $n \times p$ con entradas

$$[(AB)C]_{ij} = \sum_{l=1}^{k} (AB)_{il} c_{lj} = \sum_{l=1}^{k} \left(\sum_{s=1}^{m} a_{is} b_{sl} \right) c_{lj} = \sum_{l=1}^{k} \sum_{s=1}^{m} a_{is} b_{sl} c_{lj}.$$

Al intercambiar el orden de la suma del lado derecho nos da

$$[(AB)C]_{ij} = \sum_{s=1}^{m}\sum_{l=1}^{k} a_{is}b_{sl}c_{lj} = [A(BC)]_{ij},$$

para cada $i = 1, 2, \ldots, n$ y $j = 1, 2, \ldots, p$. Por lo que $A(BC) = (AB)C$. ■

Matrices cuadradas

Las matrices que tienen el mismo número de filas como columnas son especialmente importantes en aplicaciones.

Definición 6.9

i) Una matriz cuadrada tiene el mismo número de filas que de columnas.

ii) Una matriz diagonal $D = [d_{ij}]$ es una matriz cuadrada con $d_{ij} = 0$ siempre que $i \neq j$.

iii) La matriz identidad de orden n, $I_n = [\delta_{ij}]$, es una matriz diagonal cuyas entradas diagonales son todas 1. Cuando el tamaño de I_n es claro, en general, la matriz se escribe simplemente como I. ■

El término *diagonal* aplicado a una matriz se refiere a las entradas en la diagonal que van desde la entrada superior izquierda hasta la entrada inferior derecha.

Por ejemplo, la matriz identidad de orden 3 es

$$I = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Definición 6.10 Una matriz $n \times n$ **triangular superior** $U = [u_{ij}]$ tiene, para cada $j = 1, 2, \ldots, n$, las entradas

$$u_{ij} = 0, \quad \text{para cada } i = j+1, j+2, \ldots, n;$$

y una matriz **triangular inferior** $L = [l_{ij}]$ tiene, para cada $j = 1, 2, \ldots, n$ las entradas

$$l_{ij} = 0, \quad \text{para cada } i = 1, 2, \ldots, j-1.$$ ■

Una matriz triangular es aquella que tiene todas las entradas cero, excepto ya sea encima (superior) o debajo (inferior) de la diagonal principal.

Una matriz diagonal, entonces, es tanto triangular superior como triangular inferior debido a que sus entradas diferentes de cero deben estar en la diagonal principal.

Ilustración Considere la matriz identidad de orden 3,

$$I_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Si A es cualquier matriz 3×3, entonces

$$AI_3 = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} = A.$$ ■

La matriz identidad I_n conmuta con cualquier matriz $n \times n$ A; es decir, el orden de multiplicación no importa,

$$I_n A = A = AI_n.$$

Considere que esta propiedad, en general, no es cierta, incluso para las matrices cuadradas.

Matrices inversas

La **inversa de una matriz** está relacionada con los sistemas lineales.

Definición 6.11 Se dice que una matriz A $n \times n$ **es no singular** (o *invertible*) si existe una matriz A^{-1} $n \times n$ con $AA^{-1} = A^{-1}A = I$. La matriz A^{-1} recibe el nombre de **inversa** de A. Una matriz que no tenga inversa recibe el nombre de **singular** (o *no invertible*). ■

La palabra "singular" significa que algo se desvía de lo ordinario. Por lo tanto, una matriz singular no tiene inversa.

Las siguientes propiedades respecto a las inversas surgen a partir de la definición 6.11. Las pruebas de estos resultados se consideran en el ejercicio 13.

Teorema 6.12 Para cualquier matriz $n \times n$ no singular A:

i) A^{-1} es única.

ii) A^{-1} es no singular y $(A^{-1})^{-1} = A$.

iii) Si B también es una matriz no singular $n \times n$, entonces $(AB)^{-1} = B^{-1}A^{-1}$. ■

Ejemplo 4 Si

$$A = \begin{bmatrix} 1 & 2 & -1 \\ 2 & 1 & 0 \\ -1 & 1 & 2 \end{bmatrix} \quad \text{y} \quad B = \begin{bmatrix} -\frac{2}{9} & \frac{5}{9} & -\frac{1}{9} \\ \frac{4}{9} & -\frac{1}{9} & \frac{2}{9} \\ -\frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{bmatrix}.$$

Muestre que $B = A^{-1}$ y que la solución para el sistema lineal descrito mediante

$$\begin{aligned} x_1 + 2x_2 - x_3 &= 2, \\ 2x_1 + x_2 \phantom{{}+2x_3} &= 3, \\ -x_1 + x_2 + 2x_3 &= 4, \end{aligned}$$

está dado por las entradas en $B\mathbf{b}$, donde $\mathbf{b}$ es el vector columna con entradas 2, 3 y 4, respectivamente.

Solución Primero note que

$$AB = \begin{bmatrix} 1 & 2 & -1 \\ 2 & 1 & 0 \\ -1 & 1 & 2 \end{bmatrix} \cdot \begin{bmatrix} -\frac{2}{9} & \frac{5}{9} & -\frac{1}{9} \\ \frac{4}{9} & -\frac{1}{9} & \frac{2}{9} \\ -\frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = I_3.$$

De forma similar, $BA = I_3$, por lo que A y B son no singulares con $B = A^{-1}$ y $A = B^{-1}$.

Ahora convierta el sistema lineal dado en una ecuación matricial

$$\begin{bmatrix} 1 & 2 & -1 \\ 2 & 1 & 0 \\ -1 & 1 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 2 \\ 3 \\ 4 \end{bmatrix}$$

y multiplique ambos lados por B, la inversa de A. Como

$$B(A\mathbf{x}) = (BA)\mathbf{x} = I_3\mathbf{x} = \mathbf{x} \quad \text{y} \quad B(A\mathbf{x}) = \mathbf{b},$$

tenemos

$$BA\mathbf{x} = \left(\begin{bmatrix} -\frac{2}{9} & \frac{5}{9} & -\frac{1}{9} \\ \frac{4}{9} & -\frac{1}{9} & \frac{2}{9} \\ -\frac{3}{9} & \frac{3}{9} & \frac{3}{9} \end{bmatrix} \begin{bmatrix} 1 & 2 & -1 \\ 2 & 1 & 0 \\ -1 & 1 & 2 \end{bmatrix} \right) \mathbf{x} = \mathbf{x}$$

y

$$BA\mathbf{x} = B(\mathbf{b}) = \begin{bmatrix} -\frac{2}{9} & \frac{5}{9} & -\frac{1}{9} \\ \frac{4}{9} & -\frac{1}{9} & \frac{2}{9} \\ -\frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{bmatrix} \begin{bmatrix} 2 \\ 3 \\ 4 \end{bmatrix} = \begin{bmatrix} \frac{7}{9} \\ \frac{13}{9} \\ \frac{5}{3} \end{bmatrix}.$$

Esto implica que $\mathbf{x} = B\mathbf{b}$ y obtenemos la solución $x_1 = 7/9, x_2 = 13/9$, y $x_3 = 5/3$. ■

Aunque es fácil resolver un sistema lineal de la forma $A\mathbf{x} = \mathbf{b}$ si A^{-1}, se conoce, no es computacionalmente eficiente determinar A^{-1} con el fin de resolver el sistema. (Consulte el ejercicio 16.) Aun así, es útil desde un punto de vista conceptual describir un método para determinar la inversa de una matriz.

Para encontrar un método para calcular A^{-1} al suponer que A es no singular, observemos nuevamente la multiplicación de matrices. Si B_j es la j-ésima columna de la matriz $n \times n$ B,

$$B_j = \begin{bmatrix} b_{1j} \\ b_{2j} \\ \vdots \\ b_{nj} \end{bmatrix}.$$

Si $AB = C$, entonces la j-ésima columna de C está dada por el producto

$$\begin{bmatrix} c_{1j} \\ c_{2j} \\ \vdots \\ c_{nj} \end{bmatrix} = C_j = AB_j = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix} \begin{bmatrix} b_{1j} \\ b_{2j} \\ \vdots \\ b_{nj} \end{bmatrix} = \begin{bmatrix} \sum_{k=1}^{n} a_{1k}b_{kj} \\ \sum_{k=1}^{n} a_{2k}b_{kj} \\ \vdots \\ \sum_{k=1}^{n} a_{nk}b_{kj} \end{bmatrix}.$$

Suponga que A^{-1} existe y que $A^{-1} = B = (b_{ij})$. Entonces $AB = I$ y

$$AB_j = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \quad \text{donde el valor 1 aparece en la } j\text{-ésima fila.}$$

Para encontrar B, necesitamos resolver n sistemas lineales en los que la j-ésima columna de la inversa es la solución del sistema lineal con el lado derecho de la j-ésima columna de I. La siguiente ilustración demuestra este método.

Ilustración Para determinar la inversa de la matriz

$$A = \begin{bmatrix} 1 & 2 & -1 \\ 2 & 1 & 0 \\ -1 & 1 & 2 \end{bmatrix},$$

consideremos primero el producto AB, donde B es una matriz arbitraria 3×3:

$$AB = \begin{bmatrix} 1 & 2 & -1 \\ 2 & 1 & 0 \\ -1 & 1 & 2 \end{bmatrix} \begin{bmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \\ b_{31} & b_{32} & b_{33} \end{bmatrix}$$

$$= \begin{bmatrix} b_{11}+2b_{21}-b_{31} & b_{12}+2b_{22}-b_{32} & b_{13}+2b_{23}-b_{33} \\ 2b_{11}+b_{21} & 2b_{12}+b_{22} & 2b_{13}+b_{23} \\ -b_{11}+b_{21}+2b_{31} & -b_{12}+b_{22}+2b_{32} & -b_{13}+b_{23}+2b_{33} \end{bmatrix}.$$

Si $B = A^{-1}$, entonces, $AB = I$, de manera que

$$\begin{aligned} b_{11}+2b_{21}-\ b_{31}&=1, \\ 2b_{11}+\ b_{21}\quad\ &=0, \\ -b_{11}+\ b_{21}+2b_{31}&=0, \end{aligned} \qquad \begin{aligned} b_{12}+2b_{22}-\ b_{32}&=0, \\ 2b_{12}+\ b_{22}\quad\ &=1, \\ -b_{12}+\ b_{22}+2b_{32}&=0, \end{aligned} \qquad \text{y} \qquad \begin{aligned} b_{13}+2b_{23}-\ b_{33}&=0, \\ 2b_{13}+\ b_{23}\quad\ &=0, \\ -b_{13}+\ b_{23}+2b_{33}&=1. \end{aligned}$$

Observe que los coeficientes en cada uno de los sistemas de ecuaciones son los mismos, el único cambio en los sistemas se presenta en el lado derecho de las ecuaciones. Como consecuencia, la eliminación gaussiana se puede realizar en una matriz aumentada formada al combinar las matrices de cada uno de los sistemas:

$$\left[\begin{array}{rrr:rrr} 1 & 2 & -1 & 1 & 0 & 0 \\ 2 & 1 & 0 & 0 & 1 & 0 \\ -1 & 1 & 2 & 0 & 0 & 1 \end{array}\right].$$

Primero, realice $(E_2 - 2E_1) \to (E_2)$ y $(E_3 + E_1) \to (E_3)$, seguido de $(E_3 + E_2) \to (E_3)$, produce

$$\left[\begin{array}{rrr:rrr} 1 & 2 & -1 & 1 & 0 & 0 \\ 0 & -3 & 2 & -2 & 1 & 0 \\ 0 & 3 & 1 & 1 & 0 & 1 \end{array}\right] \quad \text{y} \quad \left[\begin{array}{rrr:rrr} 1 & 2 & -1 & 1 & 0 & 0 \\ 0 & -3 & 2 & -2 & 1 & 0 \\ 0 & 0 & 3 & -1 & 1 & 1 \end{array}\right].$$

La sustitución hacia atrás se realiza en cada una de las tres matrices aumentadas,

$$\left[\begin{array}{rrr:r} 1 & 2 & -1 & 1 \\ 0 & -3 & 2 & -2 \\ 0 & 0 & 3 & -1 \end{array}\right], \left[\begin{array}{rrr:r} 1 & 2 & -1 & 0 \\ 0 & -3 & 2 & 1 \\ 0 & 0 & 3 & 1 \end{array}\right], \left[\begin{array}{rrr:r} 1 & 2 & -1 & 0 \\ 0 & -3 & 2 & 0 \\ 0 & 0 & 3 & 1 \end{array}\right],$$

para, al final, proporcionar

$$\begin{aligned} b_{11} &= -\tfrac{2}{9}, \\ b_{21} &= \tfrac{4}{9}, \\ b_{31} &= -\tfrac{1}{3}, \end{aligned} \qquad \begin{aligned} b_{12} &= \tfrac{5}{9}, \\ b_{22} &= -\tfrac{1}{9}, \\ b_{32} &= \tfrac{1}{3}, \end{aligned} \qquad \text{y} \qquad \begin{aligned} b_{13} &= -\tfrac{1}{9}, \\ b_{23} &= \tfrac{2}{9}, \\ b_{32} &= \tfrac{1}{3}. \end{aligned}$$

Como se muestra en el ejemplo 4, éstas son las entradas de A^{-1}:

$$B = A^{-1} = \begin{bmatrix} -\frac{2}{9} & \frac{5}{9} & -\frac{1}{9} \\ \frac{4}{9} & -\frac{1}{9} & \frac{2}{9} \\ -\frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{bmatrix}.$$ ■

Como observamos en la ilustración, con el fin de calcular A^{-1}, es conveniente configurar la matriz aumentada más grande,

$$[\, A \ \vdots \ I \,].$$

Al realizar la eliminación de acuerdo con el algoritmo 6.1, obtenemos una matriz aumentada de la forma

$$[\ U \ \vdots \ Y \ ],$$

donde U es una matriz triangular superior y Y es la matriz obtenida al realizar las mismas operaciones en la identidad I que se realizaron para llevar A a U.

La eliminación gaussiana con sustitución hacia atrás requiere

$$\frac{4}{3}n^3 - \frac{1}{3}n \text{ multiplicaciones/divisiones} \quad \text{y} \quad \frac{4}{3}n^3 - \frac{3}{2}n^2 + \frac{n}{6} \text{ sumas/restas}$$

para resolver los n sistemas lineales (consulte el ejercicio 16a). Se debe tener mucho cuidado en la implementación para observar las operaciones que *no* se deben realizar, por ejemplo, una multiplicación cuando se sabe que uno de los multiplicadores es la unidad o una resta cuando se sabe que el sustraendo es 0. El número de multiplicaciones/divisiones requeridas se puede reducir a n^3 y el número de sumas/restas se puede reducir a $n^3 - 2n^2 + n$ (consulte el ejercicio 16d).

Transpuesta de una matriz

Otra matriz importante relacionada con una matriz dada A es su *transpuesta*, denotada A^t.

Definición 6.13 La **transpuesta** de una matriz $A = [a_{ij}]$ $n \times m$ es la matriz $A^t = [a_{ij}]$, $m \times n$, donde para cada i, la i-ésima columna de A^t es la misma que la i-ésima fila de A. Una matriz cuadrada A recibe el nombre de simétrica si $A = A^t$. ■

Ilustración Las matrices

$$A = \begin{bmatrix} 7 & 2 & 0 \\ 3 & 5 & -1 \\ 0 & 5 & -6 \end{bmatrix}, \quad B = \begin{bmatrix} 2 & 4 & 7 \\ 3 & -5 & -1 \end{bmatrix}, \quad C = \begin{bmatrix} 6 & 4 & -3 \\ 4 & -2 & 0 \\ -3 & 0 & 1 \end{bmatrix}$$

tienen transpuestas

$$A^t = \begin{bmatrix} 7 & 3 & 0 \\ 2 & 5 & 5 \\ 0 & -1 & -6 \end{bmatrix}, \quad B^t = \begin{bmatrix} 2 & 3 \\ 4 & -5 \\ 7 & -1 \end{bmatrix}, \quad C^t = \begin{bmatrix} 6 & 4 & -3 \\ 4 & -2 & 0 \\ -3 & 0 & 1 \end{bmatrix}.$$

La matriz C es simétrica porque $C^t = C$. Las matrices A y B no son simétricas. ■

La prueba del siguiente resultado se sigue directamente de la definición de la transpuesta.

Teorema 6.14 Las siguientes operaciones relacionadas con la transpuesta de una matriz se mantienen siempre que la operación sea posible

i) $(A^t)^t = A$,

ii) $(A + B)^t = A^t + B^t$,

iii) $(AB)^t = B^t A^t$,

iv) si A^{-1} existe, entonces $(A^{-1})^t = (A^t)^{-1}$. ■

La sección Conjunto de ejercicios 6.3 está disponible en línea. Encuentre la ruta de acceso en las páginas preliminares.

6.4 Determinante de una matriz

El *determinante* de una matriz proporciona resultados de existencia y unicidad para sistemas lineales que tienen el mismo número de incógnitas y de ecuaciones. Nosotros denotaremos el determinante de una matriz cuadrada A mediante det$A$, pero también es común utilizar la notación $|A|$.

Definición 6.15 Suponga que A es una matriz cuadrada.

i) Si $A = [a]$ es una matriz 1×1, entonces $\det A = a$.

ii) Si A es una matriz $n \times n$, con $n > 1$, el **menor** M_{ij} es el determinante de la submatriz $(n-1) \times (n-1)$ de A obtenida al quitar la i-ésima fila y la j-ésima columna de la matriz A.

iii) El **cofactor** A_{ij} asociado con M_{ij} está definido por $A_{ij} = (-1)^{i+j} M_{ij}$.

iv) El **determinante** de la matriz A $n \times n$, cuando $n > 1$, está dado ya sea por

$$\det A = \sum_{j=1}^{n} a_{ij} A_{ij} = \sum_{j=1}^{n} (-1)^{i+j} a_{ij} M_{ij}, \quad \text{para } i = 1, 2, \dots, n,$$

o mediante

$$\det A = \sum_{i=1}^{n} a_{ij} A_{ij} = \sum_{i=1}^{n} (-1)^{i+j} a_{ij} M_{ij}, \quad \text{para cualquier } j = 1, 2, \dots, n. \quad ■$$

La noción de determinante apareció de manera independiente en 1683 tanto en Japón como en Europa, a pesar de que ni Takakazu Seki Kowa (1642–1708) ni Gottfried Leibniz (1646–1716) parecían haber usado el término "determinante".

Se puede mostrar (consulte el ejercicio 12) que calcular el determinante de una matriz general $n \times n$ mediante esta definición requiere $O(n!)$ multiplicaciones/divisiones y sumas/restas. Incluso para valores relativamente pequeños de n, el número de cálculos se vuelve difícil de manejar.

A pesar de que parece que existen $2n$ definiciones diferentes de det$A$, dependiendo de la fila o columna seleccionada, todas las definiciones llevan al mismo resultado numérico. La flexibilidad en la definición se utiliza en el siguiente ejemplo. Es más conveniente calcular det$A$ a lo largo de la fila o de la columna con la mayor cantidad de ceros.

Ejemplo 1 Encuentre el determinante de la matriz

$$A = \begin{bmatrix} 2 & -1 & 3 & 0 \\ 4 & -2 & 7 & 0 \\ -3 & -4 & 1 & 5 \\ 6 & -6 & 8 & 0 \end{bmatrix}$$

mediante la fila o columna con la mayor cantidad de entradas cero.

Solución Para calcular det$A$, es más fácil utilizar la cuarta columna:

$$\det A = a_{14}A_{14} + a_{24}A_{24} + a_{34}A_{34} + a_{44}A_{44} = 5A_{34} = -5M_{34}.$$

Al eliminar la tercera fila y la cuarta columna obtenemos

$$\det A = -5 \det \begin{bmatrix} 2 & -1 & 3 \\ 4 & -2 & 7 \\ 6 & -6 & 8 \end{bmatrix}$$

$$= -5 \left\{ 2 \det \begin{bmatrix} -2 & 7 \\ -6 & 8 \end{bmatrix} - (-1) \det \begin{bmatrix} 4 & 7 \\ 6 & 8 \end{bmatrix} + 3 \det \begin{bmatrix} 4 & -2 \\ 6 & -6 \end{bmatrix} \right\} = -30. \quad ■$$

Las siguientes propiedades son útiles para relacionar sistemas lineales y eliminación gaussiana para determinantes. Éstas están probadas en cualquier texto de álgebra lineal estándar.

Teorema 6.16 Suponga que A es una matriz $n \times n$:

i) Si cualquier fila o columna A sólo tiene entradas cero, entonces $\det A = 0$.

ii) Si A tiene dos filas o dos columnas iguales, entonces $\det A = 0$.

iii) Si $\tilde{A}$ se obtiene a partir de A mediante la operación $(E_i) \leftrightarrow (E_j)$, con $i \neq j$, entonces $\det \tilde{A} = -\det A$.

iv) Si $\tilde{A}$ se obtiene a partir de A mediante la operación $(\lambda E_i) \to (E_i)$, entonces $\tilde{A} = \lambda \det A$.

v) Si $\tilde{A}$ se obtiene a partir de A mediante la operación $(E_i + \lambda E_j) \to (E_i)$ con $i \neq j$, entonces $\det \tilde{A} = \det A$.

vi) Si B también es una matriz $n \times n$, entonces $\det AB = \det A \det B$.

vii) $\det A^t = \det A$.

viii) Cuando A^{-1} existe, $\det A^{-1} = (\det A)^{-1}$.

ix) Si A es una matriz triangular superior, triangular inferior o diagonal, entonces $A = \prod_{i=1}^{n} a_{ii}$. ■

La parte ix) del teorema 6.16 indica que el determinante de una matriz triangular es simplemente el producto de sus elementos diagonales. Al emplear las operaciones de fila dadas en las partes iii), iv) y v), podemos reducir una matriz cuadrada determinada a la forma triangular para encontrar su determinante.

Ejemplo 2 Calcule el determinante de la matriz

$$A = \begin{bmatrix} 2 & 1 & -1 & 1 \\ 1 & 1 & 0 & 3 \\ -1 & 2 & 3 & -1 \\ 3 & -1 & -1 & 2 \end{bmatrix}$$

por medio de las partes iii), iv) y v) del teorema 6.16.

Solución La secuencia de las operaciones en la tabla 6.2 produce la matriz

$$\tilde{A} = \begin{bmatrix} 1 & \frac{1}{2} & -\frac{1}{2} & \frac{1}{2} \\ 0 & 1 & 1 & 5 \\ 0 & 0 & 3 & 13 \\ 0 & 0 & 0 & -13 \end{bmatrix}.$$

Por medio de la parte ix), $\det \tilde{A} = -39$, luego $\det A = 39$. ■

Tabla 6.2

Operación	Efecto
$\frac{1}{2}E_1 \to E_1$	$\det A1 = \frac{1}{2}\det A$
$E_2 - E_1 \to E_2$	$\det A2 = \det A1 = \frac{1}{2}\det A$
$E_3 + E_1 \to E_3$	$\det A3 = \det A2 = \frac{1}{2}\det A$
$E_4 - 3E_1 \to E_4$	$\det A4 = \det A3 = \frac{1}{2}\det A$
$2E_2 \to E_2$	$\det A5 = 2\det A4 = \det A$
$E_3 - \frac{5}{2}E_2 \to E_3$	$\det A6 = \det A5 = \det A$
$E_4 + \frac{5}{2}E_2 \to E_4$	$\det A7 = \det A6 = \det A$
$E_3 \leftrightarrow E_4$	$\det A8 = -\det A7 = -\det A$

El resultado clave relacionado con no singularidad, eliminación gaussiana, sistemas lineales y determinantes es que las siguientes declaraciones son equivalentes.

Teorema 6.17 Las siguientes declaraciones son equivalentes para cualquier matriz $n \times n$ A:

i) La ecuación $A\mathbf{x} = 0$ tiene la solución única $\mathbf{x} = \mathbf{0}$.

ii) El sistema $A\mathbf{x} = \mathbf{b}$ tiene una solución única para cualquier vector de columna n-dimensional $\mathbf{b}$.

iii) La matriz A es no singular, es decir, existe A^{-1}.

iv) $\det A \neq 0$.

v) La eliminación gaussiana con intercambios de fila se puede realizar en el sistema $A\mathbf{x} = \mathbf{b}$ para cualquier vector de columna n-dimensional $\mathbf{b}$. ■

El siguiente corolario para el teorema 6.17 ilustra cómo se puede utilizar el determinante para mostrar propiedades importantes sobre matrices cuadradas.

Corolario 6.18 Suponga que tanto A como B son matrices $n \times n$ ya sea con $AB = I$ o $BA = I$. Entonces $B = A^{-1}$ (y $A = B^{-1}$).

Demostración Suponga que $AB = I$. Entonces, por medio de la parte vi) del teorema 6.16

$$1 = \det(I) = \det(AB) = \det(A) \cdot \det(B), \quad \text{por lo que} \quad \det(A) \neq 0 \text{ y } \det(B) \neq 0.$$

La equivalencia de las partes iii) y iv) del teorema 6.17 implica que existe tanto A^{-1} como B^{-1}. Por lo tanto,

$$A^{-1} = A^{-1} \cdot I = A^{-1} \cdot (AB) = \left(A^{-1}A\right) \cdot B = I \cdot B = B.$$

Los papeles de A y B son similares, por lo que esto también establece que $BA = I$. Por lo que $B = A^{-1}$. ■

La sección Conjunto de ejercicios 6.4 está disponible en línea. Encuentre la ruta de acceso en las páginas preliminares.

6.5 Factorización de matriz

La eliminación gaussiana es la herramienta principal en la solución directa de sistemas de ecuaciones lineales, por lo que no debería ser sorprendente que aparezca de otras formas. En esta sección, observaremos que los pasos para resolver un sistema de la forma $A\mathbf{x} = \mathbf{b}$ se pueden usar para factorizar una matriz. La factorización es especialmente útil cuando tiene la forma $A = LU$, donde L es triangular inferior y U es triangular superior. A pesar de que no todas las matrices tienen este tipo de representación, muchas se presentan con frecuencia en la aplicación de técnicas numéricas.

En la sección 6.1 encontramos que la eliminación gaussiana aplicada a un sistema lineal arbitrario $A\mathbf{x} = \mathbf{b}$ requiere $O(n^3/3)$ operaciones aritméticas para determinar $\mathbf{x}$. Sin embargo, resolver un sistema lineal que implica un sistema triangular superior sólo requiere sustitución hacia atrás, que toma $O(n^2)$ operaciones. El número de operaciones requeridas para resolver un sistema triangular inferior es similar.

Suponga que A se ha factorizado en la forma triangular $A = LU$, donde L es triangular inferior y U es triangular superior. Entonces podemos resolver para $\mathbf{x}$ con mayor facilidad a través de un proceso de dos pasos.

- Primero, hacemos $\mathbf{y} = U\mathbf{x}$ y resolvemos el sistema triangular superior $L\mathbf{y} = \mathbf{b}$ para y. Puesto que L es triangular, determinar $\mathbf{y}$ a partir de esta ecuación sólo requiere $O(n^2)$ operaciones.

- Una vez que conocemos $\mathbf{y}$, el sistema triangular superior $U\mathbf{x} = \mathbf{y}$ solamente requiere una operación $O(n^2)$ adicional para determinar la solución de $\mathbf{x}$.

La resolución de un sistema lineal $A\mathbf{x} = \mathbf{b}$ en forma factorizada significa que el número de operaciones necesario para resolver el sistema $A\mathbf{x} = \mathbf{b}$ se reduce a partir de $O(n^3/3)$ a $O(2n^2)$.

Ejemplo 1 Compare el número aproximado de operaciones requeridas para determinar la solución para un sistema lineal mediante una técnica que requiera $O(n^3/3)$ operaciones y una que necesite $O(2n^2)$ cuando $n = 20$, $n = 100$ y $n = 1000$.

Solución La tabla 6.3 muestra los resultados de estos cálculos. ■

Tabla 6.3

n	$n^3/3$	$2n^2$	% de reducción
10	$3.\overline{3} \times 10^2$	2×10^2	40
100	$3.\overline{3} \times 10^5$	2×10^4	94
1000	$3.\overline{3} \times 10^8$	2×10^6	99.4

Como ilustra el ejemplo, el factor de reducción aumenta drásticamente con el tamaño de la matriz. No es sorprendente que las reducciones a partir de la factorización tengan un costo; la determinación de las matrices específicas L y U requiere $O(n^3/3)$ operaciones. Pero, una vez que se determina la factorización, los sistemas relacionados con la matriz A se pueden resolver de esta forma simplificada para cualquier número de vectores $\mathbf{b}$.

Para observar qué matrices tienen factorización LU y encontrar cómo se determina, primero suponga que la eliminación gaussiana se puede realizar en el sistema $A\mathbf{x} = \mathbf{b}$ sin intercambios de fila. Con la notación en la sección 6.1, esto es equivalente a tener $a_{ii}^{(i)}$, elementos pivote diferentes de cero, para cada $i = 1, 2, \ldots, n$.

El primer paso en el proceso de eliminación gaussiana consiste en realizar, para cada $j = 2, 3, \ldots, n$, las operaciones

$$(E_j - m_{j,1}E_1) \to (E_j), \quad \text{donde} \quad m_{j,1} = \frac{a_{j1}^{(1)}}{a_{11}^{(1)}}. \tag{6.8}$$

Estas operaciones transforman el sistema en uno en el que todas las entradas en la primera columna por debajo de la diagonal son cero.

El sistema de operaciones en la ecuación (6.8) se puede observar de otra manera. Se logra simultáneamente al multiplicar la matriz original A a la izquierda de la matriz

$$M^{(1)} = \begin{bmatrix} 1 & 0 & \cdots & \cdots & 0 \\ -m_{21} & 1 & \ddots & & \vdots \\ \vdots & 0 & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ -m_{n1} & 0 & \cdots & 0 & 1 \end{bmatrix}.$$

La factorización de matrices es otra de las técnicas importantes que Gauss parece haber descubierto primero. Está incluida en su tratado de dos volúmenes sobre mecánica celeste *Theoria motus corporum coelestium in sectionibus conicis Solem ambientium*, publicado en 1809.

Esto recibe el nombre de **primera matriz de transformación gaussiana**. Nosotros denotamos el producto de esta matriz con $A^{(1)} \equiv A$ por $A^{(2)}$ y con $\mathbf{b}$ y $\mathbf{b}^{(2)}$, por lo que

$$A^{(2)}\mathbf{x} = M^{(1)}A\mathbf{x} = M^{(1)}\mathbf{b} = \mathbf{b}^{(2)}.$$

De manera similar, construimos $M^{(2)}$, la matriz identidad con entradas por debajo de la diagonal en la segunda columna reemplazadas por los negativos de los multiplicadores

$$m_{j,2} = \frac{a_{j2}^{(2)}}{a_{22}^{(2)}}.$$

El producto de esta matriz con $A^{(2)}$ tiene ceros por debajo de la diagonal en las primeras dos columnas y hacemos

$$A^{(3)}\mathbf{x} = M^{(2)}A^{(2)}\mathbf{x} = M^{(2)}M^{(1)}A\mathbf{x} = M^{(2)}M^{(1)}\mathbf{b} = \mathbf{b}^{(3)}.$$

En general, con $A^{(k)}\mathbf{x} = \mathbf{b}^{(k)}$ ya formada, multiplicamos por la ***k*-ésima matriz de transformación gaussiana**

$$M^{(k)} = \begin{bmatrix} 1 & 0 & \cdots & \cdots & \cdots & \cdots & \cdots & 0 \\ 0 & \ddots & \ddots & & & & & \vdots \\ \vdots & \ddots & \ddots & \ddots & & & & \vdots \\ \vdots & & 0 & \ddots & \ddots & & & \vdots \\ \vdots & & \vdots & -m_{k+1,k} & \ddots & \ddots & & \vdots \\ \vdots & & \vdots & \vdots & 0 & \ddots & \ddots & \vdots \\ \vdots & & \vdots & \vdots & \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & -m_{n,k} & 0 & \cdots & 0 & 1 \end{bmatrix},$$

para obtener

$$A^{(k+1)}\mathbf{x} = M^{(k)}A^{(k)}\mathbf{x} = M^{(k)}\cdots M^{(1)}A\mathbf{x} = M^{(k)}\mathbf{b}^{(k)} = \mathbf{b}^{(k+1)} = M^{(k)}\cdots M^{(1)}\mathbf{b}. \quad (6.9)$$

El proceso termina con la formación de $A^{(n)}\mathbf{x} = \mathbf{b}^{(n)}$, donde $A^{(n)}$, es la matriz triangular superior

$$A^{(n)} = \begin{bmatrix} a_{11}^{(1)} & a_{12}^{(1)} & \cdots & a_{1n}^{(1)} \\ 0 & a_{22}^{(2)} & \ddots & \vdots \\ \vdots & \ddots & \ddots & a_{n-1,n}^{(n-1)} \\ 0 & \cdots & 0 & a_{nn}^{(n)} \end{bmatrix},$$

dada por

$$A^{(n)} = M^{(n-1)}M^{(n-2)}\cdots M^{(1)}A.$$

Este proceso forma la parte $U = A^{(n)}$ de la factorización de la matriz $A = LU$. Para determinar la matriz L triangular inferior complementaria, primero recuerde la multiplicación de $A^{(k)}\mathbf{x} = \mathbf{b}^{(k)}$ por la transformación gaussiana de $\boldsymbol{M^{(k)}}$ utilizada para obtener la ecuación (6.9):

$$A^{(k+1)}\mathbf{x} = M^{(k)}A^{(k)}\mathbf{x} = M^{(k)}\mathbf{b}^{(k)} = \mathbf{b}^{(k+1)},$$

donde $M^{(k)}$ genera las operaciones de la fila

$$(E_j - m_{j,k}E_k) \to (E_j), \quad \text{para } j = k+1, \ldots, n.$$

Invertir los efectos de esta transformación y regresar a $A^{(k)}$ requiere realizar las operaciones $(E_j + m_{j,k}E_k) \rightarrow (E_j)$ para cada $j = k+1, \ldots, n$. esto es equivalente a multiplicar por la inversa de la matriz $\boldsymbol{M}^{(k)}$, la matriz

$$L^{(k)} = [M^{(k)}]^{-1} = \begin{bmatrix} 1 & 0 & \cdots & \cdots & \cdots & \cdots & 0 \\ 0 & \ddots & \ddots & & & & \vdots \\ \vdots & \ddots & 0 & \ddots & & & \vdots \\ \vdots & & \vdots & m_{k+1,k} & \ddots & & \vdots \\ \vdots & & \vdots & \vdots & 0 & \ddots & \vdots \\ \vdots & & \vdots & \vdots & \vdots & \ddots & 0 \\ 0 & \cdots & 0 & m_{n,k} & 0 & \cdots 0 & 1 \end{bmatrix}.$$

La matriz L triangular inferior en la factorización de A, entonces, es el producto de las matrices $L^{(k)}$:

$$L = L^{(1)}L^{(2)} \cdots L^{(n-1)} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ m_{21} & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ m_{n1} & \cdots & m_{n,n-1} & 1 \end{bmatrix},$$

puesto que el producto de L con la matriz triangular superior $U = M^{(n-1)} \cdots M^{(2)}M^{(1)}A$ da

$$\begin{aligned} LU &= L^{(1)}L^{(2)} \cdots L^{(n-3)}L^{(n-2)}L^{(n-1)} \cdot M^{(n-1)}M^{(n-2)}M^{(n-3)} \cdots M^{(2)}M^{(1)}A \\ &= [M^{(1)}]^{-1}[M^{(2)}]^{-1} \cdots [M^{(n-2)}]^{-1}[M^{(n-1)}]^{-1} \cdot M^{(n-1)}M^{(n-2)} \cdots M^{(2)}M^{(1)}A = A. \end{aligned}$$

El teorema 6.19 sigue estas observaciones.

Teorema 6.19 Si la eliminación gaussiana se puede realizar en el sistema lineal $A\mathbf{x} = \mathbf{b}$ sin intercambios de fila, entonces la matriz A se puede factorizar en el producto de una matriz triangular inferior L y una matriz triangular superior U, es decir, $A = LU$, donde $m_{ji} = a_{ji}^{(i)}/a_{ii}^{(i)}$,

$$U = \begin{bmatrix} a_{11}^{(1)} & a_{12}^{(1)} & \cdots & a_{1n}^{(1)} \\ 0 & a_{22}^{(2)} & \ddots & \vdots \\ \vdots & \ddots & \ddots & a_{n-1,n}^{(n-1)} \\ 0 & \cdots & 0 & a_{nn}^{(n)} \end{bmatrix}, \quad \text{y} \quad L = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ m_{21} & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ m_{n1} & \cdots & m_{n,n-1} & 1 \end{bmatrix}.$$

Ejemplo 2 **a)** Determine la factorización LU para la matriz A en el sistema lineal $A\mathbf{x} = \mathbf{b}$, donde

$$A = \begin{bmatrix} 1 & 1 & 0 & 3 \\ 2 & 1 & -1 & 1 \\ 3 & -1 & -1 & 2 \\ -1 & 2 & 3 & -1 \end{bmatrix} \quad \text{y} \quad \mathbf{b} = \begin{bmatrix} 4 \\ 1 \\ -3 \\ 4 \end{bmatrix}.$$

b) Entonces utilice la factorización para resolver el sistema

$$\begin{aligned} x_1 + x_2 \quad\quad + 3x_4 &= 8, \\ 2x_1 + x_2 - x_3 + x_4 &= 7, \\ 3x_1 - x_2 - x_3 + 2x_4 &= 14, \\ -x_1 + 2x_2 + 3x_3 - x_4 &= -7. \end{aligned}$$

Solución **a)** El sistema original se consideró en la sección 6.1, donde observamos que la secuencia de operaciones $(E_2 - 2E_1) \to (E_2), (E_3 - 3E_1) \to (E_3), (E_4 - (-1)E_1) \to (E_4), (E_3 - 4E_2) \to (E_3), (E_4 - (-3)E_2) \to (E_4)$ convierte el sistema en el sistema triangular

$$\begin{aligned} x_1 + x_2 \quad\quad + \; 3x_4 &= \;\;\, 4, \\ - x_2 - \;\; x_3 - \;\; 5x_4 &= -7, \\ 3x_3 + 13x_4 &= \;\;\, 13, \\ - 13x_4 &= -13. \end{aligned}$$

Los multiplicadores m_{ij} y la matriz triangular superior producen la factorización

$$A = \begin{bmatrix} 1 & 1 & 0 & 3 \\ 2 & 1 & -1 & 1 \\ 3 & -1 & -1 & 2 \\ -1 & 2 & 3 & -1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ 3 & 4 & 1 & 0 \\ -1 & -3 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 & 0 & 3 \\ 0 & -1 & -1 & -5 \\ 0 & 0 & 3 & 13 \\ 0 & 0 & 0 & -13 \end{bmatrix} = LU.$$

b) Para resolver

$$A\mathbf{x} = LU\mathbf{x} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ 3 & 4 & 1 & 0 \\ -1 & -3 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 & 0 & 3 \\ 0 & -1 & -1 & -5 \\ 0 & 0 & 3 & 13 \\ 0 & 0 & 0 & -13 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 8 \\ 7 \\ 14 \\ -7 \end{bmatrix},$$

primero introducimos la sustitución $\mathbf{y} = U\mathbf{x}$. Entonces $\mathbf{b} = L(U\mathbf{x}) = L\mathbf{y}$. Es decir,

$$L\mathbf{y} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ 3 & 4 & 1 & 0 \\ -1 & -3 & 0 & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} = \begin{bmatrix} 8 \\ 7 \\ 14 \\ -7 \end{bmatrix}.$$

Este sistema se resuelve para $\mathbf{y}$ mediante un proceso de sustitución hacia adelante:

$$\begin{aligned} y_1 &= 8; \\ 2y_1 + y_2 &= 7, && \text{por lo que} \quad y_2 = 7 - 2y_1 = -9; \\ 3y_1 + 4y_2 + y_3 &= 14, && \text{por lo que} \quad y_3 = 14 - 3y_1 - 4y_2 = 26; \\ -y_1 - 3y_2 + y_4 &= -7, && \text{por lo que} \quad y_4 = -7 + y_1 + 3y_2 = -26. \end{aligned}$$

Entonces, resolvemos $U\mathbf{x} = \mathbf{y}$ para $\mathbf{x}$, la solución del sistema original; es decir,

$$\begin{bmatrix} 1 & 1 & 0 & 3 \\ 0 & -1 & -1 & -5 \\ 0 & 0 & 3 & 13 \\ 0 & 0 & 0 & -13 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 8 \\ -9 \\ 26 \\ -26 \end{bmatrix}.$$

Utilizando la sustitución hacia atrás, obtenemos $x_4 = 2, x_3 = 0, x_2 = -1, x_1 = 3$. ■

La factorización utilizada en el ejemplo 2 recibe el nombre de *método de Doolittle* y requiere que los 1 estén en la diagonal de L, lo cual resulta en la factorización descrita en el teorema 6.19. En la sección 6.6, consideramos el *método de Crout*, factorización que requiere que los elementos en la diagonal de U sean 1, y el *método de Choleski*, que requiere que $l_{ii} = u_{ii}$, para cada i.

Un procedimiento general para factorización de matrices como un producto de matrices triangulares se encuentra en el algoritmo 6.4. A pesar de que se construyen matrices nuevas L *y* U, los valores generados pueden reemplazar las entradas correspondientes de A que ya no son necesarias.

El algoritmo 6.4 permite especificar la diagonal de L o la diagonal de U.

ALGORITMO 6.4

Factorización *LU*

Para factorizar la matriz $n \times n$ $A = [a_{ij}]$ en el producto de la matriz triangular inferior $L = [l_{ij}]$ y la matriz triangular superior $U = [u_{ij}]$, es decir $A = LU$, donde la diagonal principal ya sea de L o U consta sólo de unos:

ENTRADA dimensión n; las entradas a_{ij}, $1 \leq i, j \leq n$ de A; la diagonal $l_{11} = \cdots = l_{nn} = 1$ de L o la diagonal $u_{11} = \cdots = u_{nn} = 1$ de U.

SALIDA las entradas l_{ij}, $1 \leq j \leq i$, $1 \leq i \leq n$ de L y las entradas u_{ij}, $i \leq j \leq n$, $1 \leq i \leq n$ de U.

Paso 1 Seleccione l_{11} y u_{11} al satisfacer $l_{11}u_{11} = a_{11}$.
Si $l_{11}u_{11} = 0$ entonces SALIDA ('Factorización imposible');
PARE.

Paso 2 Para $j = 2, \ldots, n$ determine $u_{1j} = a_{1j}/l_{11}$; (*Primera fila de U.*)
$l_{j1} = a_{j1}/u_{11}$. (*Primera columna de L.*)

Paso 3 Para $i = 2, \ldots, n-1$ haga los pasos 4 y 5.

Paso 4 Seleccione l_{ii} y u_{ii} al satisfacer $l_{ii}u_{ii} = a_{ii} - \sum_{k=1}^{i-1} l_{ik}u_{ki}$.
Si $l_{ii}u_{ii} = 0$ entonces SALIDA ('Factorización imposible');
PARE.

Paso 5 Para $j = i+1, \ldots, n$

Determine $u_{ij} = \frac{1}{l_{ii}}\left[a_{ij} - \sum_{k=1}^{i-1} l_{ik}u_{kj}\right]$; (*i-ésima fila de U.*)

$l_{ji} = \frac{1}{u_{ii}}\left[a_{ji} - \sum_{k=1}^{i-1} l_{jk}u_{ki}\right]$. (*i-ésima columna de L.*)

Paso 6 Seleccione l_{nn} y u_{nn} al satisfacer $l_{nn}u_{nn} = a_{nn} - \sum_{k=1}^{n-1} l_{nk}u_{kn}$.
(*Nota: Si* $l_{nn}u_{nn} = 0$, *entonces* $A = LU$ *pero A es singular.*)

Paso 7 SALIDA (l_{ij} para $j = 1, \ldots, i$ y $i = 1, \ldots, n$);
SALIDA (u_{ij} para $j = i, \ldots, n$ y $i = 1, \ldots, n$);
PARE. ■

Una vez que se completa la factorización de la matriz, la solución de un sistema lineal de la forma $A\mathbf{x} = LU\mathbf{x} = \mathbf{b}$ se obtiene primero, haciendo $\mathbf{y} = U\mathbf{x}$ y resolver $L\mathbf{y} = \mathbf{b}$ para y. Puesto que L es triangular inferior, tenemos

$$y_1 = \frac{b_1}{l_{11}},$$

y, para cada $i = 2, 3, \ldots, n$,

$$y_i = \frac{1}{l_{ii}}\left[b_i - \sum_{j=1}^{i-1} l_{ij}y_j\right].$$

Después de encontrar $\mathbf{y}$ mediante este proceso de sustitución hacia adelante, el sistema triangular superior $U\mathbf{x} = \mathbf{y}$ se resuelve para $\mathbf{x}$ mediante sustitución hacia atrás con las ecuaciones

$$x_n = \frac{y_n}{u_{nn}} \quad \text{y} \quad x_i = \frac{1}{u_{ii}}\left[y_i - \sum_{j=i+1}^{n} u_{ij}x_j\right].$$

Matrices de permutación

En el análisis previo suponemos que $A\mathbf{x} = \mathbf{b}$ se puede resolver por medio de eliminación gaussiana sin intercambios de fila. Desde un punto de vista práctico, esta factorización es útil sólo cuando no se requieren intercambios de fila para controlar el error de redondeo que resulta a partir del uso de aritmética de dígitos finitos. Por fortuna muchos sistemas que encontramos al utilizar métodos de aproximación son de este tipo, pero ahora consideraremos las modificaciones que se deben hacer cuando se requieren cambios de fila. Comenzamos el análisis con la introducción de una clase de matrices que se usan para reordenar, o permutar, filas de una matriz determinada.

Una **matriz** $n \times n$ **de permutación** $P = [p_{ij}]$ es una matriz obtenida al reordenar las filas de I_n, la matriz identidad. Esto produce una matriz con precisamente una entrada diferente a cero en cada fila y en cada columna, y cada entrada diferente a cero es un 1.

Ilustración La matriz

$$P = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}$$

es una matriz de permutación 3×3. Para cualquier matriz A 3×3, multiplicar a la izquierda por P tiene el efecto de intercambiar la segunda y tercera filas de A:

$$PA = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{31} & a_{32} & a_{33} \\ a_{21} & a_{22} & a_{23} \end{bmatrix}.$$

De igual forma, multiplicar A a la derecha por P intercambia la segunda y tercera columnas de A. ■

Dos propiedades útiles de las matrices de permutación se relacionan con la eliminación gaussiana, la primera de las cuales se ilustra en el ejemplo previo. Suponga que $k_1, \ldots, k_n$ es una permutación de los enteros $1, \ldots, n$ y la matriz de permutación $P = (p_{ij})$ se define mediante

$$p_{ij} = \begin{cases} 1, & \text{si } j = k_i, \\ 0, & \text{en otro caso.} \end{cases}$$

Entonces

La multiplicación de la matriz AP permuta las columnas de A.

- PA permuta las filas de A; es decir,

$$PA = \begin{bmatrix} a_{k_1 1} & a_{k_1 2} & \cdots & a_{k_1 n} \\ a_{k_2 1} & a_{k_2 2} & \cdots & a_{k_2 n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{k_n 1} & a_{k_n 2} & \cdots & a_{k_n n} \end{bmatrix}.$$

- P^{-1} existe y $P^{-1} = P^t$.

Al final de la sección 6.4, observamos que para cualquier matriz no singular A, el sistema lineal $A\mathbf{x} = \mathbf{b}$ se puede resolver mediante la eliminación gaussiana, sin la posibilidad de intercambios de filas. Si conocemos los intercambios de filas requeridos para resolver el sistema con eliminación gaussiana, podemos acomodar las ecuaciones originales en un orden que garantizaría que no se necesiten intercambios de fila. Por lo tanto, existe una reorganización de ecuaciones en el sistema que permite eliminación gaussiana para proceder

sin intercambios de fila. Esto implica que para una matriz no singular A, existe una matriz de permutación P para la que el sistema

$$PA\mathbf{x} = P\mathbf{b}$$

se puede resolver sin intercambios de fila. Como consecuencia, esta matriz PA se puede factorizar en

$$PA = LU,$$

donde L es triangular inferior y U es triangular superior. Puesto que $P^{-1} = P^t$, esto produce la factorización

$$A = P^{-1}LU = (P^tL)U.$$

La matriz U sigue siendo triangular superior, pero P^tL no es triangular inferior a menos que $P = I$.

Ejemplo 3 Determine una factorización en la forma $A = (P^tL)U$ para la matriz

$$A = \begin{bmatrix} 0 & 0 & -1 & 1 \\ 1 & 1 & -1 & 2 \\ -1 & -1 & 2 & 0 \\ 1 & 2 & 0 & 2 \end{bmatrix}.$$

Solución La matriz A no puede tener una factorización LU porque $a_{11} = 0$. Sin embargo, utilizar un intercambio de fila $(E_1) \leftrightarrow (E_2)$, seguido por $(E_3 + E_1) \rightarrow (E_3)$ y $(E_4 - E_1) \rightarrow (E_4)$, produce

$$\begin{bmatrix} 1 & 1 & -1 & 2 \\ 0 & 0 & -1 & 1 \\ 0 & 0 & 1 & 2 \\ 0 & 1 & 1 & 0 \end{bmatrix}.$$

Entonces el intercambio de fila $(E_2) \leftrightarrow (E_4)$, seguido de $(E_4 + E_3) \rightarrow (E_4)$, produce la matriz

$$U = \begin{bmatrix} 1 & 1 & -1 & 2 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 2 \\ 0 & 0 & 0 & 3 \end{bmatrix}.$$

La matriz de permutación relacionada con los intercambios de fila $(E_1) \leftrightarrow (E_2)$ y $(E_2) \leftrightarrow (E_4)$ es

$$P = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix}$$

y

$$PA = \begin{bmatrix} 1 & 1 & -1 & 2 \\ 1 & 2 & 0 & 2 \\ -1 & -1 & 2 & 0 \\ 0 & 0 & -1 & 1 \end{bmatrix}.$$

La eliminación gaussiana se realiza en PA mediante las mismas operaciones que en A, excepto sin intercambios de fila. Es decir, $(E_2 - E_1) \rightarrow (E_2)$, $(E_3 + E_1) \rightarrow (E_3)$, seguido de $(E_4 + E_3) \rightarrow (E_4)$. Los multiplicadores diferentes a cero para PA son, por consiguiente,

$$m_{21} = 1, \quad m_{31} = -1, \quad \text{y} \quad m_{43} = -1,$$

y la factorización LU de PA es

$$PA = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ -1 & 0 & 1 & 0 \\ 0 & 0 & -1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 & -1 & 2 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 2 \\ 0 & 0 & 0 & 3 \end{bmatrix} = LU.$$

Multiplicar por $P^{-1} = P^t$ produce la factorización

$$A = P^{-1}(LU) = P^t(LU) = (P^tL)U = \begin{bmatrix} 0 & 0 & -1 & 1 \\ 1 & 0 & 0 & 0 \\ -1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & 1 & -1 & 2 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 2 \\ 0 & 0 & 0 & 3 \end{bmatrix}. \quad ■$$

La sección Conjunto de ejercicios 6.5 está disponible en línea. Encuentre la ruta de acceso en las páginas preliminares.

6.6 Tipos especiales de matrices

Ahora nos enfocaremos en dos clases de matrices para las que la eliminación gaussiana se puede realizar de manera efectiva sin intercambios de fila.

Matrices diagonalmente dominantes

La primera clase se describe en la siguiente definición.

Definición 6.20 Se dice que la matriz A $n \times n$ es **diagonalmente dominante** cuando

$$|a_{ii}| \geq \sum_{\substack{j=1,\\ j\neq i}}^{n} |a_{ij}| \quad \text{se mantiene para cada } i = 1, 2, \ldots, n. \tag{6.10}$$

Cada entrada de la diagonal principal en una matriz estrictamente diagonalmente dominante tiene una magnitud que es estrictamente superior a la suma de las magnitudes de todas las otras entradas en esa fila.

Una matriz diagonalmente dominante es **estrictamente diagonalmente dominante** cuando la desigualdad en la ecuación (6.10) es estricta para cada n, es decir, cuando

$$|a_{ii}| > \sum_{\substack{j=1,\\ j\neq i}}^{n} |a_{ij}| \quad \text{se mantiene para cada } i = 1, 2, \ldots, n. \quad ■$$

Ilustración Considere las matrices

$$A = \begin{bmatrix} 7 & 2 & 0 \\ 3 & 5 & -1 \\ 0 & 5 & -6 \end{bmatrix} \quad \text{y} \quad B = \begin{bmatrix} 6 & 4 & -3 \\ 4 & -2 & 0 \\ -3 & 0 & 1 \end{bmatrix}.$$

La matriz no simétrica A es estrictamente diagonalmente dominante porque

$$|7| > |2| + |0|, \quad |5| > |3| + |-1|, \quad \text{y} \quad |-6| > |0| + |5|.$$

La matriz simétrica B no es estrictamente diagonalmente dominante porque, por ejemplo, en la primera fila, el valor absoluto del elemento diagonal es $|6| < |4| + |-3| = 7$. Es interesante observar que A^t no es estrictamente diagonalmente dominante porque la fila del medio de A^t es [2 5 5], ni, por supuesto es B^t porque $B^t = B$. ■

El siguiente teorema se utilizó en la sección 3.5 para garantizar que existen soluciones únicas para los sistemas lineales necesarios para determinar spline cúbicos interpolantes.

Teorema 6.21 Una matriz estrictamente diagonalmente dominante A es no singular. Además, en este caso, la eliminación gaussiana se puede realizar en cualquier sistema lineal de la forma $A\mathbf{x} = \mathbf{b}$ para obtener su única solución sin intercambios de fila o columna y los cálculos serán estables respecto al crecimiento de errores de redondeo.

Demostración Primero aplicamos la demostración por contradicción para mostrar que A es no singular. Considere el sistema lineal descrito por $A\mathbf{x} = \mathbf{0}$ y suponga que existe una solución distinta a cero $\mathbf{x} = (x_i)$ para este sistema. Si k es un índice para el que

$$0 < |x_k| = \max_{1 \leq j \leq n} |x_j|.$$

Puesto que $\sum_{j=1}^{n} a_{ij} x_j = 0$ para cada $i = 1, 2, \ldots, n$, tenemos, cuando $i = k$,

$$a_{kk} x_k = -\sum_{\substack{j=1, \\ j \neq k}}^{n} a_{kj} x_j.$$

A partir de la desigualdad triangular, tenemos

$$|a_{kk}||x_k| \leq \sum_{\substack{j=1, \\ j \neq k}}^{n} |a_{kj}||x_j|, \quad \text{por lo que} \quad |a_{kk}| \leq \sum_{\substack{j=1, \\ j \neq k}}^{n} |a_{kj}| \frac{|x_j|}{|x_k|} \leq \sum_{\substack{j=1, \\ j \neq k}}^{n} |a_{kj}|.$$

Esta desigualdad contradice la dominancia diagonal estricta de A. Por consiguiente, la única solución para $A\mathbf{x} = \mathbf{0}$ es $\mathbf{x} = \mathbf{0}$. Esto muestra, en el teorema 6.17 en la página 298, ser equivalente a la no singularidad de A.

Para probar que la eliminación gaussiana se puede realizar sin intercambios de fila, mostramos que cada una de las matrices $A^{(2)}, A^{(3)}, \ldots, A^{(n)}$ generadas por el proceso de eliminación gaussiana (y descrito en la sección 6.5) es estrictamente diagonalmente dominante. Eso garantizaría que en cada etapa del proceso de eliminación gaussiana, el elemento pivote es distinto a cero.

Puesto que A es estrictamente diagonalmente dominante, $a_{11} \neq 0$ y $A^{(2)}$ se puede realizar. Por lo tanto, para cada $i = 2, 3, \ldots, n$,

$$a_{ij}^{(2)} = a_{ij}^{(1)} - \frac{a_{1j}^{(1)} a_{i1}^{(1)}}{a_{11}^{(1)}}, \quad \text{para} \quad 2 \leq j \leq n.$$

Primero, $a_{i1}^{(2)} = 0$. La desigualdad triangular implica que

$$\sum_{\substack{j=2 \\ j \neq i}}^{n} |a_{ij}^{(2)}| = \sum_{\substack{j=2 \\ j \neq i}}^{n} \left| a_{ij}^{(1)} - \frac{a_{1j}^{(1)} a_{i1}^{(1)}}{a_{11}^{(1)}} \right| \leq \sum_{\substack{j=2 \\ j \neq i}}^{n} |a_{ij}^{(1)}| + \sum_{\substack{j=2 \\ j \neq i}}^{n} \left| \frac{a_{1j}^{(1)} a_{i1}^{(1)}}{a_{11}^{(1)}} \right|.$$

Pero puesto que A es estrictamente diagonalmente dominante,

$$\sum_{\substack{j=2\\ j\neq i}}^{n} |a_{ij}^{(1)}| < |a_{ii}^{(1)}| - |a_{i1}^{(1)}| \quad \text{y} \quad \sum_{\substack{j=2\\ j\neq i}}^{n} |a_{1j}^{(1)}| < |a_{11}^{(1)}| - |a_{1i}^{(1)}|,$$

por lo que

$$\sum_{\substack{j=2\\ j\neq i}}^{n} |a_{ij}^{(2)}| < |a_{ii}^{(1)}| - |a_{i1}^{(1)}| + \frac{|a_{i1}^{(1)}|}{|a_{11}^{(1)}|}(|a_{11}^{(1)}| - |a_{1i}^{(1)}|) = |a_{ii}^{(1)}| - \frac{|a_{i1}^{(1)}||a_{1i}^{(1)}|}{|a_{11}^{(1)}|}.$$

La desigualdad triangular también implica que

$$|a_{ii}^{(1)}| - \frac{|a_{i1}^{(1)}||a_{1i}^{(1)}|}{|a_{11}^{(1)}|} \leq \left| a_{ii}^{(1)} - \frac{|a_{i1}^{(1)}||a_{1i}^{(1)}|}{|a_{11}^{(1)}|} \right| = |a_{ii}^{(2)}|,$$

lo cual nos da

$$\sum_{\substack{j=2\\ j\neq i}}^{n} |a_{ij}^{(2)}| < |a_{ii}^{(2)}|.$$

Esto establece el dominio diagonal estricto para las filas $2, \ldots, n$. Pero la primera fila de $A^{(2)}$ y A son la misma, por lo que $A^{(2)}$ es estrictamente diagonalmente dominante.

Este proceso continúa de manera inductiva hasta que se obtiene $A^{(n)}$ triangular superior y estrictamente diagonalmente dominante. Esto implica que todos los elementos diagonales son diferentes a cero, por lo que se puede realizar la eliminación gaussiana sin intercambios de fila.

La demostración de estabilidad para este procedimiento se puede encontrar en [We]. ■

Matrices definidas positivas

La siguiente clase especial de matrices recibe el nombre de *definida positiva*.

Definición 6.22 Una matriz A es **definida positiva** si es simétrica y si $\mathbf{x}^t A\mathbf{x} > 0$ para cada vector n-dimensional $\mathbf{x} \neq \mathbf{0}$. ■

El nombre definida positiva se refiere al hecho de que el número $\mathbf{x}^t A\mathbf{x}$ debe ser positivo siempre que $\mathbf{x} \neq \mathbf{0}$.

No todos los autores requieren simetría de una matriz definida positiva. Por ejemplo, Golub y van Loan (GV), una referencia estándar en métodos matriciales sólo requieren que $\mathbf{x}^t A\mathbf{x} > 0$ para cada $\mathbf{x} \neq 0$. Las matrices que llamamos definidas positivas reciben el nombre de definida positiva simétrica en [GV]. Mantenga esta discrepancia en mente si utiliza material de otras fuentes.

Para ser preciso, la definición 6.22 debería especificar que la matriz 1×1 generada por la operación $\mathbf{x}^t A\mathbf{x}$ tiene un valor positivo para su única entrada ya que la operación se realiza de acuerdo con lo siguiente:

$$\mathbf{x}^t A\mathbf{x} = [x_1, x_2, \ldots, x_n] \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

$$= [x_1, x_2, \ldots, x_n] \begin{bmatrix} \sum_{j=1}^{n} a_{1j}x_j \\ \sum_{j=1}^{n} a_{2j}x_j \\ \vdots \\ \sum_{j=1}^{n} a_{nj}x_j \end{bmatrix} = \left[\sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij}x_i x_j \right].$$

Ejemplo 1 Muestre que la matriz

$$A = \begin{bmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{bmatrix}$$

es definida positiva.

Solución Suponga que $\mathbf{x}$ es cualquier vector columna tridimensional. Entonces

$$\begin{aligned} \mathbf{x}^t A\mathbf{x} &= [x_1, x_2, x_3] \begin{bmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \\ &= [x_1, x_2, x_3] \begin{bmatrix} 2x_1 - x_2 \\ -x_1 + 2x_2 - x_3 \\ -x_2 + 2x_3 \end{bmatrix} \\ &= 2x_1^2 - 2x_1x_2 + 2x_2^2 - 2x_2x_3 + 2x_3^2. \end{aligned}$$

Al reorganizar los términos obtenemos

$$\begin{aligned} \mathbf{x}^t A\mathbf{x} &= x_1^2 + (x_1^2 - 2x_1x_2 + x_2^2) + (x_2^2 - 2x_2x_3 + x_3^2) + x_3^2 \\ &= x_1^2 + (x_1 - x_2)^2 + (x_2 - x_3)^2 + x_3^2, \end{aligned}$$

lo cual implica que

$$x_1^2 + (x_1 - x_2)^2 + (x_2 - x_3)^2 + x_3^2 > 0$$

a menos que $x_1 = x_2 = x_3 = 0$. ■

A partir del ejemplo debería quedar claro que utilizar la definición para determinar si una matriz es definida positiva puede ser difícil. Afortunadamente, existen criterios de verificación más fáciles, presentados en el capítulo 9, para identificar a los miembros de esta importante clase. El siguiente resultado proporciona algunas condiciones necesarias que se pueden usar para determinar ciertas matrices de la consideración.

Teorema 6.23 Si A es una matriz $n \times n$ definida positiva, entonces

i) A tiene una inversa;

ii) $a_{ii} > 0$, para cada $i = 1, 2, \ldots, n$;

iii) $\max_{1\le k,j\le n} |a_{kj}| \le \max_{1\le i\le n} |a_{ii}|$;

iv) $(a_{ij})^2 < a_{ii}a_{jj}$ para cada $i \neq j$.

Demostración

i) Si $\mathbf{x}$ satisface $A\mathbf{x} = \mathbf{0}$, entonces $\mathrm{x}^t A\mathbf{x} = \mathbf{0}$. Ya que A es definida positiva, esto implica $\mathbf{x} = \mathbf{0}$. Por consiguiente, $A\mathbf{x} = \mathbf{0}$ sólo tiene la solución cero. Mediante el teorema 6.17 en la página 298, esto es equivalente a que A no sea singular.

ii) Para una i determinada, si $\mathbf{x} = (x_j)$ se puede definir mediante $x_i = 1$ y $x_j = 0$, si $j \neq i$. Puesto que $\mathbf{x} \neq \mathbf{0}$,

$$0 < \mathbf{x}^t A\mathbf{x} = a_{ii}.$$

iii) Para $k \neq j$, defina $\mathbf{x} = (x_i)$ mediante

$$x_i = \begin{cases} 0, & \text{si } i \neq j \text{ y } i \neq k, \\ 1, & \text{si } i = j, \\ -1, & \text{si } i = k. \end{cases}$$

Puesto que $\mathbf{x} \neq \mathbf{0}$,

$$0 < \mathbf{x}^t A\mathbf{x} = a_{jj} + a_{kk} - a_{jk} - a_{kj}.$$

Pero $A^t = A$, por lo que $a_{jk} = a_{kj}$, lo cual implica que

$$2a_{kj} < a_{jj} + a_{kk}. \tag{6.11}$$

Ahora defina $\mathbf{z} = (z_i)$ mediante

$$z_i = \begin{cases} 0, & \text{si } i \neq j \text{ y } i \neq k, \\ 1, & \text{si } i = j \text{ o } i = k. \end{cases}$$

Entonces $\mathbf{z}^t A\mathbf{z} > 0$, por lo que

$$-2a_{kj} < a_{kk} + a_{jj}. \tag{6.12}$$

Las ecuaciones (6.11) y (6.12) implican que para cada $k \neq j$,

$$|a_{kj}| < \frac{a_{kk} + a_{jj}}{2} \leq \max_{1 \leq i \leq n} |a_{ii}|, \quad \text{por lo que} \quad \max_{1 \leq k, j \leq n} |a_{kj}| \leq \max_{1 \leq i \leq n} |a_{ii}|.$$

iv) Para $i \neq j$, defina $\mathbf{x} = (x_k)$ mediante

$$x_k = \begin{cases} 0, & \text{si } k \neq j \text{ y } k \neq i, \\ \alpha, & \text{si } k = i, \\ 1, & \text{si } k = j, \end{cases}$$

donde α representa un número real arbitrario. Puesto que $\mathbf{x} \neq \mathbf{0}$,

$$0 < \mathbf{x}^t A\mathbf{x} = a_{ii}\alpha^2 + 2a_{ij}\alpha + a_{jj}.$$

Como polinomio cuadrático en α sin raíces reales, el discriminante de $P(\alpha) = a_{ii}\alpha^2 + 2a_{ij}\alpha + a_{jj}$ debe ser negativo. Por lo tanto,

$$4a_{ij}^2 - 4a_{ii}a_{jj} < 0 \quad \text{y} \quad a_{ij}^2 < a_{ii}a_{jj}.$$ ■

A pesar de que el teorema 6.23 provee algunas condiciones importantes que deben ser verdaderas para matrices definidas positivas, no garantiza que una matriz que satisfaga estas condiciones sea definida positiva.

La siguiente noción se usará para proporcionar una condición necesaria y suficiente.

Definición 6.24 Una **primera submatriz principal** de una matriz A es una matriz de la forma

$$A_k = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1k} \\ a_{21} & a_{22} & \cdots & a_{2k} \\ \vdots & \vdots & & \vdots \\ a_{k1} & a_{k2} & \cdots & a_{kk} \end{bmatrix},$$

para algunas $1 \leq k \leq n$. ■

Una demostración del siguiente resultado se puede encontrar en [Stew1], p. 250.

Teorema 6.25 Una matriz simétrica A es definida positiva si y sólo si cada una de sus primeras submatrices principales tiene un determinante positivo. ■

Ejemplo 2 En el ejemplo 1 utilizamos la definición para mostrar que la matriz simétrica

$$A = \begin{bmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{bmatrix}$$

es definida positiva. Confirme esto con el teorema 6.25.

Solución Observe que

$$\det A_1 = \det[2] = 2 > 0,$$

$$\det A_2 = \det \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix} = 4 - 1 = 3 > 0,$$

y

$$\det A_3 = \det \begin{bmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{bmatrix} = 2 \det \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix} - (-1) \det \begin{bmatrix} -1 & -1 \\ 0 & 2 \end{bmatrix}$$
$$= 2(4 - 1) + (-2 + 0) = 4 > 0,$$

de acuerdo con el teorema 6.25. ■

El siguiente resultado amplía la parte **i)** del teorema 6.23 e iguala los resultados estrictamente diagonalmente dominantes presentados en el teorema 6.21 en la página 307. No probaremos este teorema porque requiere introducir terminología y resultados que no son necesarios para cualquier otro propósito. El desarrollo y la demostración se pueden encontrar en [We], p. 120 ff.

Teorema 6.26 Una matriz simétrica A es una definida positiva si y sólo si se puede realizar eliminación gaussiana sin intercambios de fila en el sistema lineal $A\mathbf{x} = \mathbf{b}$ con todos los elementos pivote positivos. Además, en este caso, los cálculos son estables respecto al crecimiento de errores de redondeo. ■

Algunos hechos interesantes no cubiertos al construir la demostración del teorema 6.26 se presentan en los siguientes corolarios.

Corolario 6.27 La matriz A es definida positiva si y sólo si A se puede factorizar en la forma LDL^t, donde L es triangular inferior con números 1 en su diagonal y D es una matriz diagonal con entradas diagonales positivas. ■

Corolario 6.28 La matriz A es definida positiva si y sólo si A se puede factorizar en la forma LL^t, donde L es triangular inferior con entradas diagonales diferentes a cero. ■

La matriz L en el corolario 6.28 no es igual a la matriz L en el corolario 6.27. Una relación entre ellas se presenta en el ejercicio 32.

El algoritmo 6.25 está basado en el algoritmo de factorización LU 6.4 y obtiene la factorización LDL^t descrita en el corolario 6.27.

ALGORITMO **6.5**

Factorización LDL^t

Para factorizar la matriz A $n \times n$ definida positiva en la forma LDL^t, donde L es una matriz triangular inferior con 1 a lo largo de la diagonal y D es una matriz diagonal con entradas positivas en la diagonal:

ENTRADA la dimensión n; entradas a_{ij}, para $1 \le i, j \le n$ de A.

SALIDA las entradas l_{ij}, para $1 \le j < i$ y $1 \le i \le n$ de L, y d_i, para $1 \le i \le n$ de D.

Paso 1 Para $i = 1, \ldots, n$ haga los pasos 2–4.

Paso 2 Para $j = 1, \ldots, i - 1$, determine $v_j = l_{ij} d_j$.



Paso 3 Determine $d_i = a_{ii} - \sum_{j=1}^{i-1} l_{ij} v_j$.

Paso 4 Para $j = i+1, \ldots, n$ determine $l_{ji} = (a_{ji} - \sum_{k=1}^{i-1} l_{jk} v_k)/d_i$.

Paso 5 SALIDA (l_{ij} para $j = 1, \ldots, i-1$ y $i = 1, \ldots, n$);
SALIDA (d_i para $i = 1, \ldots, n$);
PARE. ■

El corolario 6.27 tiene una contraparte cuando A es simétrica, pero no necesariamente definida positiva. Este resultado se aplica de manera amplia porque las matrices simétricas son comunes y se reorganizan con facilidad.

Corolario 6.29 Si A es una matriz simétrica $n \times n$ para el que se puede aplicar eliminación gaussiana sin intercambios de fila. Entonces A se puede factorizar en LDL^t, donde L es triangular inferior con números 1 en su diagonal y D es la matriz diagonal con $a_{11}^{(1)}, \ldots, a_{nn}^{(n)}$ en su diagonal. ■

Ejemplo 3 Determine la factorización LDL^t de la matriz definida positiva.

$$A = \begin{bmatrix} 4 & -1 & 1 \\ -1 & 4.25 & 2.75 \\ 1 & 2.75 & 3.5 \end{bmatrix}.$$

Solución La factorización LDL^t tiene números 1 en la diagonal de la matriz triangular inferior L, por lo que necesitamos tener

$$A = \begin{bmatrix} a_{11} & a_{21} & a_{31} \\ a_{21} & a_{22} & a_{32} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ l_{21} & 1 & 0 \\ l_{31} & l_{32} & 1 \end{bmatrix} \begin{bmatrix} d_1 & 0 & 0 \\ 0 & d_2 & 0 \\ 0 & 0 & d_3 \end{bmatrix} \begin{bmatrix} 1 & l_{21} & l_{31} \\ 0 & 1 & l_{32} \\ 0 & 0 & 1 \end{bmatrix}$$

$$= \begin{bmatrix} d_1 & d_1 l_{21} & d_1 l_{31} \\ d_1 l_{21} & d_2 + d_1 l_{21}^2 & d_2 l_{32} + d_1 l_{21} l_{31} \\ d_1 l_{31} & d_1 l_{21} l_{31} + d_2 l_{32} & d_1 l_{31}^2 + d_2 l_{32}^2 + d_3 \end{bmatrix}.$$

Por lo tanto,

$a_{11}: 4 = d_1 \implies d_1 = 4,$ $\quad a_{21}: -1 = d_1 l_{21} \implies l_{21} = -0.25$

$a_{31}: 1 = d_1 l_{31} \implies l_{31} = 0.25,$ $\quad a_{22}: 4.25 = d_2 + d_1 l_{21}^2 \implies d_2 = 4$

$a_{32}: 2.75 = d_1 l_{21} l_{31} + d_2 l_{32} \implies l_{32} = 0.75,$ $\quad a_{33}: 3.5 = d_1 l_{31}^2 + d_2 l_{32}^2 + d_3 \implies d_3 = 1,$

y tenemos

$$A = LDL^t = \begin{bmatrix} 1 & 0 & 0 \\ -0.25 & 1 & 0 \\ 0.25 & 0.75 & 1 \end{bmatrix} \begin{bmatrix} 4 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & -0.25 & 0.25 \\ 0 & 1 & 0.75 \\ 0 & 0 & 1 \end{bmatrix}.$$ ■

André-Louis Cholesky (1875–1918) fue un oficial militar francés involucrado en la geodesia y la topografía a principios de 1900. Él desarrolló este método de factorización para calcular soluciones a los problemas de mínimos cuadrados.

El algoritmo 6.5 se modifica fácilmente para factorizar las matrices simétricas descritas en el corolario 6.29. Simplemente requiere añadir una verificación para garantizar que los elementos diagonales sean diferentes a cero. El algoritmo de Cholesky 6.6 produce la factorización LL^t, descrita en el corolario 6.28.

ALGORITMO
6.6

Factorización de Cholesky

Para factorizar la matriz definida positiva A $n \times n$ en LL^t, donde L es triangular inferior:

ENTRADA la dimensión n; entradas a_{ij}, para $\leq i, j \leq n$ de A.

SALIDA las entradas l_{ij}, para $1 \leq j \leq i$ y $1 \leq i \leq n$ de L. (*Las entradas de* $U = L^t$ *son* $u_{ij} = l_{ji}$, *para* $i \leq j \leq$ *y* $1 \leq i \leq n$.)

Paso 1 Determine $l_{11} = \sqrt{a_{11}}$.

Paso 2 Para $j = 2, \ldots, n$, determine $l_{j1} = a_{j1}/l_{11}$.

Paso 3 Para $i = 2, \ldots, n-1$ haga los pasos 4 y 5.

Paso 4 Determine $l_{ii} = \left(a_{ii} - \sum_{k=1}^{i-1} l_{ik}^2\right)^{1/2}$.

Paso 5 Para $j = i+1, \ldots, n$

determine $l_{ji} = \left(a_{ji} - \sum_{k=1}^{i-1} l_{jk}l_{ik}\right)/l_{ii}$.

Paso 6 Determine $l_{nn} = \left(a_{nn} - \sum_{k=1}^{n-1} l_{nk}^2\right)^{1/2}$.

Paso 7 SALIDA (l_{ij} para $j = 1, \ldots, i$ y $i = 1, \ldots, n$);
PARE. ■

Ejemplo 4 Determine la factorización LL^t de Cholesky de la matriz definida positiva

$$A = \begin{bmatrix} 4 & -1 & 1 \\ -1 & 4.25 & 2.75 \\ 1 & 2.75 & 3.5 \end{bmatrix}.$$

Solución La factorización LL^t no necesariamente tiene números 1 en la diagonal de la matriz triangular inferior L, por lo que necesitamos tener

$$A = \begin{bmatrix} a_{11} & a_{21} & a_{31} \\ a_{21} & a_{22} & a_{32} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} = \begin{bmatrix} l_{11} & 0 & 0 \\ l_{21} & l_{22} & 0 \\ l_{31} & l_{32} & l_{33} \end{bmatrix} \begin{bmatrix} l_{11} & l_{21} & l_{31} \\ 0 & l_{22} & l_{32} \\ 0 & 0 & l_{33} \end{bmatrix}$$

$$= \begin{bmatrix} l_{11}^2 & l_{11}l_{21} & l_{11}l_{31} \\ l_{11}l_{21} & l_{21}^2 + l_{22}^2 & l_{21}l_{31} + l_{22}l_{32} \\ l_{11}l_{31} & l_{21}l_{31} + l_{22}l_{32} & l_{31}^2 + l_{32}^2 + l_{33}^2 \end{bmatrix}.$$

Por lo tanto,

$$a_{11}: \quad 4 = l_{11}^2 \implies l_{11} = 2, \qquad a_{21}: \quad -1 = l_{11}l_{21} \implies l_{21} = -0.5$$

$$a_{31}: \quad 1 = l_{11}l_{31} \implies l_{31} = 0.5, \qquad a_{22}: \quad 4.25 = l_{21}^2 + l_{22}^2 \implies l_{22} = 2$$

$$a_{32}: \quad 2.75 = l_{21}l_{31} + l_{22}l_{32} \implies l_{32} = 1.5, \qquad a_{33}: \quad 3.5 = l_{31}^2 + l_{32}^2 + l_{33}^2 \implies l_{33} = 1,$$

y tenemos

$$A = LL^t = \begin{bmatrix} 2 & 0 & 0 \\ -0.5 & 2 & 0 \\ 0.5 & 1.5 & 1 \end{bmatrix} \begin{bmatrix} 2 & -0.5 & 0.5 \\ 0 & 2 & 1.5 \\ 0 & 0 & 1 \end{bmatrix}.$$ ■

La factorización LDL^t descrita en el algoritmo 6.5 requiere

$$\frac{1}{6}n^3 + n^2 - \frac{7}{6}n \text{ multiplicaciones/divisiones} \quad \text{y} \quad \frac{1}{6}n^3 - \frac{1}{6}n \text{ sumas/restas.}$$

La factorización LL^t de Cholesky de una matriz definida positiva sólo requiere

$$\frac{1}{6}n^3 + \frac{1}{2}n^2 - \frac{2}{3}n \text{ multiplicaciones/divisiones} \quad \text{y} \quad \frac{1}{6}n^3 - \frac{1}{6}n \text{ sumas/restas.}$$

Esta ventaja computacional de la factorización de Cholesky es engañosa porque requiere extraer n raíces cuadradas. Sin embargo, el número de operaciones que se necesitan para calcular las n raíces cuadradas es un factor lineal de n y disminuirá su importancia conforme n aumenta.

El algoritmo 6.5 provee un método estable para factorizar una matriz definida positiva de la forma $A = LDL^t$, pero se debe modificar para resolver el sistema lineal $A\mathbf{x} = \mathbf{b}$. Para hacerlo, borramos la declaración PARE del paso 5 en el algoritmo y añadimos los siguientes pasos para resolver el sistema triangular inferior $L\mathbf{y} = \mathbf{b}$:

Paso 6 Determine $y_1 = b_1$.

Paso 7 Para $i = 2, \ldots, n$ determine $y_i = b_i - \sum_{j=1}^{i-1} l_{ij} y_j$.

El sistema lineal $D\mathbf{z} = \mathbf{y}$ se puede resolver por medio de

Paso 8 Para $i = 1, \ldots, n$ determine $z_i = y_i/d_i$.

Finalmente, el sistema triangular superior $L^t\mathbf{x} = \mathbf{z}$ se resuelve con los pasos dados por

Paso 9 Determine $x_n = z_n$.

Paso 10 Para $i = n - 1, \ldots,$ determine $x_i = z_i - \sum_{j=i+1}^{n} l_{ji} x_j$.

Paso 11 SALIDA (x_i para $i = 1, \ldots, n$);
PARE.

La tabla 6.4 muestra las operaciones adicionales requeridas para resolver el sistema lineal.

Tabla 6.4

Paso	Multiplicaciones/divisiones	Sumas/restas
6	0	0
7	$n(n-1)/2$	$n(n-1)/2$
8	n	0
9	0	0
10	$n(n-1)/2$	$n(n-1)/2$
Total	n^2	$n^2 - n$

Si se prefiere la factorización de Cholesky dada en el algoritmo 6.6, los pasos adicionales para resolver el sistema $A\mathbf{x} = \mathbf{b}$ son los siguientes. Primero borre la instrucción PARE del paso 7. A continuación, añada:

Paso 8 Determine $y_1 = b_1/l_{11}$.

Paso 9 Para $i = 2, \ldots, n$ determine $y_i = \left(b_i - \sum_{j=1}^{i-1} l_{ij} y_j\right) \Big/ l_{ii}$.

Paso 10 Determine $x_n = y_n/l_{nn}$.

Paso 11 Para $i = n - 1, \ldots,$ determine $x_i = \left(y_i - \sum_{j=i+1}^{n} l_{ji} x_j\right) \Big/ l_{ii}$.

Paso 12 SALIDA (x_i para $i = 1, \ldots, n$);
PARE.

Los pasos 8-12 requieren $n^2 + n$ multiplicaciones/divisiones y $n^2 - n$ sumas/restas.

Matrices de banda

La última clase de matrices considerada son las *matrices de banda*. En muchas aplicaciones, las matrices de banda también son estrictamente diagonalmente dominantes o definidas positivas.

Definición 6.30 Una matriz $n \times n$ recibe el nombre de **matriz de banda** si existen los enteros p y q con $1 < p, q < n$, con la propiedad de que $a_{ij} = 0$ siempre que $p \leq j - i$ o $q \leq i - j$. El **ancho de banda** o banda de una matriz se define como $w = p + q - 1$. ■

El nombre para una matriz de banda proviene del hecho de que todas las entradas diferentes a cero se encuentran en una banda centrada en la diagonal principal.

El número p describe el número de diagonales sobre la diagonal principal, incluyendo la diagonal principal, en la que pueden encontrar entradas diferentes a cero. El número q describe el número de diagonales debajo de la diagonal principal, incluyendo la diagonal principal, en la que pueden encontrar entradas diferentes a cero. Por ejemplo, la matriz

$$A = \begin{bmatrix} 7 & 2 & 0 \\ 3 & 5 & -1 \\ 0 & -5 & -6 \end{bmatrix}$$

es una matriz de banda con $p = q = 2$ y ancho de banda $2 + 2 - 1 = 3$.

La definición de matriz de banda obliga a estas matrices a concentrar todas sus entradas diferentes a cero alrededor de la diagonal. Dos casos especiales de matrices de banda que se presentan con frecuencia tienen $p = q = 2$ y $p = q = 4$.

Matrices tridiagonales

Las matrices de ancho de banda 3 se presentan cuando $p = q = 2$, reciben el nombre de **tridiagonales** porque tienen la forma

$$A = \begin{bmatrix} a_{11} & a_{12} & 0 & \cdots & \cdots & 0 \\ a_{21} & a_{22} & a_{23} & \ddots & & \vdots \\ 0 & a_{32} & a_{33} & a_{34} & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & a_{n-1,n} \\ 0 & \cdots & 0 & & a_{n,n-1} & a_{nn} \end{bmatrix}.$$

Las matrices tridiagonales también se consideran en el capítulo 11 junto con el estudio de las aproximaciones lineales por tramos para problemas de valor en la frontera. El caso $p = q = 4$, se utilizará para solucionar problemas de valor en la frontera al aproximar funciones que asumen la forma de splines cúbicos.

Los algoritmos de factorización se pueden simplificar considerablemente en el caso de matrices de banda debido al gran número de ceros que aparecen en estas matrices en patrones regulares. Es especialmente interesante observar la forma que el método Crout o Doolittle asume en este caso.

Para ilustrar la situación, suponga que una matriz tridiagonal A se puede factorizar en las matrices triangulares L y U. Entonces A tiene máximo $(3n - 2)$ entradas diferentes a cero. Por lo que existen solamente $(3n - 2)$ condiciones a aplicar para determinar las entradas de L y U, siempre y cuando, por supuesto, también se obtengan las entradas cero de A.

Suponga que las matrices L y U también tienen forma tridiagonal; es decir,

$$L = \begin{bmatrix} l_{11} & 0 & \cdots & \cdots & 0 \\ l_{21} & l_{22} & \ddots & & \vdots \\ 0 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & l_{n,n-1} & l_{nn} \end{bmatrix} \quad \text{y} \quad U = \begin{bmatrix} 1 & u_{12} & 0 & \cdots & 0 \\ 0 & 1 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & \ddots & u_{n-1,n} \\ 0 & \cdots & \cdots & 0 & 1 \end{bmatrix}.$$

Hay $(2n - 1)$ entradas indeterminadas de L y $(n - 1)$ entradas determinadas de U, que suman $(3n - 2)$, el número de posibles entradas diferentes a cero de A. Las entradas 0 de A se obtienen automáticamente.

La multiplicación relacionada con $A = LU$ nos da, además de las entradas 0,

$$a_{11} = l_{11};$$

$$a_{i,i-1} = l_{i,i-1}, \quad \text{para cada } i = 2, 3, \ldots, n; \tag{6.13}$$

$$a_{ii} = l_{i,i-1}u_{i-1,i} + l_{ii}, \quad \text{para cada } i = 2, 3, \ldots, n; \tag{6.14}$$

y

$$a_{i,i+1} = l_{ii}u_{i,i+1}, \quad \text{para cada } i = 1, 2, \ldots, n-1. \tag{6.15}$$

Una solución para este sistema se encuentra al utilizar primero la ecuación (6.13) para obtener los términos fuera de la diagonal diferentes a cero en L y después, con las ecuaciones (6.14) y (6.15) para obtener de manera alternativa el resto de las entradas en U y L. Una vez que se calcula una entrada L o U, la entrada correspondiente en A no es necesaria. Por lo que, las entradas en A se pueden sobrescribir mediante las entradas en L y U con el resultado de que no se requiere almacenamiento nuevo.

El algoritmo 6.7 resuelve un sistema $n \times n$ de ecuaciones lineales cuya matriz de coeficiente es tridiagonal. Este algoritmo solamente requiere $(5n - 4)$ multiplicaciones/divisiones y $(3n - 3)$ sumas/restas. Por consiguiente, tiene una ventaja computacional considerable sobre los métodos que no consideran la tridiagonalidad de la matriz.

ALGORITMO **6.7**

Factorización de Crout para sistemas lineales tridiagonales

Para resolver el sistema lineal $n \times n$

$$\begin{array}{llr}
E_1: & a_{11}x_1 + a_{12}x_2 & = a_{1,n+1}, \\
E_2: & a_{21}x_1 + a_{22}x_2 + a_{23}x_3 & = a_{2,n+1}, \\
\vdots & \qquad\qquad\vdots & \vdots \\
E_{n-1}: & a_{n-1,n-2}x_{n-2} + a_{n-1,n-1}x_{n-1} + a_{n-1,n}x_n & = a_{n-1,n+1}, \\
E_n: & a_{n,n-1}x_{n-1} + a_{nn}x_n & = a_{n,n+1},
\end{array}$$

que se supone que tiene una solución única:

ENTRADA la dimensión n; las entradas de A.

SALIDA la solución $x_1, \ldots, x_n$.

(*Los pasos 1-3 configuran y resuelven* $L\mathbf{z} = \mathbf{b}$.)

Paso 1 Determine $l_{11} = a_{11}$;
$u_{12} = a_{12}/l_{11}$;
$z_1 = a_{1,n+1}/l_{11}$.

Paso 2 Para $i = 2, \ldots, n-1$ determine $l_{i,i-1} = a_{i,i-1}$; (*i-ésima fila de L.*)
$l_{ii} = a_{ii} - l_{i,i-1}u_{i-1,i}$;
$u_{i,i+1} = a_{i,i+1}/l_{ii}$; (*(i + 1)-ésima columna de U.*)
$z_i = (a_{i,n+1} - l_{i,i-1}z_{i-1})/l_{ii}$.

Paso 3 Determine $l_{n,n-1} = a_{n,n-1}$; (*enésima fila de L.*)
$l_{nn} = a_{nn} - l_{n,n-1}u_{n-1,n}$.
$z_n = (a_{n,n+1} - l_{n,n-1}z_{n-1})/l_{nn}$.

(*Pasos 4 y 5 resuelven* $U\mathbf{x} = \mathbf{z}$.)

Paso 4 Determine $x_n = z_n$.

Paso 5 Para $i = n-1, \ldots, 1$ determine $x_i = z_i - u_{i,i+1}x_{i+1}$.

Paso 6 SALIDA $(x_1, \ldots, x_n)$;
PARE. ■

Ejemplo 5 Determine la factorización de Crout de la matriz tridiagonal simétrica

$$\begin{bmatrix} 2 & -1 & 0 & 0 \\ -1 & 2 & -1 & 0 \\ 0 & -1 & 2 & -1 \\ 0 & 0 & -1 & 2 \end{bmatrix}$$

y utilice esta factorización para resolver el sistema lineal

$$\begin{aligned} 2x_1 - x_2 \phantom{{}+2x_3 - x_4} &= 1, \\ -x_1 + 2x_2 - x_3 \phantom{{}-x_4} &= 0, \\ -x_2 + 2x_3 - x_4 &= 0, \\ -x_3 + 2x_4 &= 1. \end{aligned}$$

Solución La factorización LU de A tiene la forma

$$A = \begin{bmatrix} a_{11} & 0 & 0 & 0 \\ a_{21} & a_{22} & a_{23} & 0 \\ 0 & a_{32} & a_{33} & a_{34} \\ 0 & 0 & a_{43} & a_{44} \end{bmatrix} = \begin{bmatrix} l_{11} & 0 & 0 & 0 \\ l_{21} & l_{22} & 0 & 0 \\ 0 & l_{32} & l_{33} & 0 \\ 0 & 0 & l_{43} & l_{44} \end{bmatrix} \begin{bmatrix} 1 & u_{12} & 0 & 0 \\ 0 & 1 & u_{23} & 0 \\ 0 & 0 & 1 & u_{34} \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$$= \begin{bmatrix} l_{11} & l_{11}u_{12} & 0 & 0 \\ l_{21} & l_{22} + l_{21}u_{12} & l_{22}u_{23} & 0 \\ 0 & l_{32} & l_{33} + l_{32}u_{23} & l_{33}u_{34} \\ 0 & 0 & l_{43} & l_{44} + l_{43}u_{34} \end{bmatrix}.$$

Por lo tanto,

$$\begin{aligned} a_{11}&: \quad 2 = l_{11} \implies l_{11} = 2, & a_{12}&: \quad -1 = l_{11}u_{12} \implies u_{12} = -\tfrac{1}{2}, \\ a_{21}&: \quad -1 = l_{21} \implies l_{21} = -1, & a_{22}&: \quad 2 = l_{22} + l_{21}u_{12} \implies l_{22} = \tfrac{3}{2}, \\ a_{23}&: \quad -1 = l_{22}u_{23} \implies u_{23} = -\tfrac{2}{3}, & a_{32}&: \quad -1 = l_{32} \implies l_{32} = -1, \\ a_{33}&: \quad 2 = l_{33} + l_{32}u_{23} \implies l_{33} = \tfrac{4}{3}, & a_{34}&: \quad -1 = l_{33}u_{34} \implies u_{34} = -\tfrac{3}{4}, \\ a_{43}&: \quad -1 = l_{43} \implies l_{43} = -1, & a_{44}&: \quad 2 = l_{44} + l_{43}u_{34} \implies l_{44} = \tfrac{5}{4}. \end{aligned}$$

Esto nos da la factorización de Crout

$$A = \begin{bmatrix} 2 & -1 & 0 & 0 \\ -1 & 2 & -1 & 0 \\ 0 & -1 & 2 & -1 \\ 0 & 0 & -1 & 2 \end{bmatrix} = \begin{bmatrix} 2 & 0 & 0 & 0 \\ -1 & \frac{3}{2} & 0 & 0 \\ 0 & -1 & \frac{4}{3} & 0 \\ 0 & 0 & -1 & \frac{5}{4} \end{bmatrix} \begin{bmatrix} 1 & -\frac{1}{2} & 0 & 0 \\ 0 & 1 & -\frac{2}{3} & 0 \\ 0 & 0 & 1 & -\frac{3}{4} \\ 0 & 0 & 0 & 1 \end{bmatrix} = LU.$$

Al resolver el sistema

$$L\mathbf{z} = \begin{bmatrix} 2 & 0 & 0 & 0 \\ -1 & \frac{3}{2} & 0 & 0 \\ 0 & -1 & \frac{4}{3} & 0 \\ 0 & 0 & -1 & \frac{5}{4} \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \\ z_3 \\ z_4 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 1 \end{bmatrix} \quad \text{nos da} \quad \begin{bmatrix} z_1 \\ z_2 \\ z_3 \\ z_4 \end{bmatrix} = \begin{bmatrix} \frac{1}{2} \\ \frac{1}{3} \\ \frac{1}{4} \\ 1 \end{bmatrix},$$

y al resolver

$$U\mathbf{x} = \begin{bmatrix} 1 & -\frac{1}{2} & 0 & 0 \\ 0 & 1 & -\frac{2}{3} & 0 \\ 0 & 0 & 1 & -\frac{3}{4} \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} \frac{1}{2} \\ \frac{1}{3} \\ \frac{1}{4} \\ 1 \end{bmatrix} \quad \text{nos da} \quad \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}. \quad \blacksquare$$

El algoritmo de factorización Crout se puede aplicar siempre que $l_{ii} \neq 0$ para cada $i = 1, 2, \ldots, n$. Dos condiciones, cualquiera que garantice que esto es verdad, son que la matriz de coeficientes del sistema es definida positiva o que es estrictamente diagonalmente dominante. Una condición adicional que garantiza la aplicación de este algoritmo se proporciona en el siguiente teorema, cuya demostración se considera en el ejercicio 30.

Teorema 6.31 Suponga que $A = [a_{ij}]$ es tridiagonal con $a_{i,i-1}a_{i,i+1} \neq 0$, para cada $i = 2, 3, \ldots, n-1$. Si $|a_{11}| > |a_{12}|$, $|a_{ii}| \geq |a_{i,i-1}| + |a_{i,i+1}|$, para cada $i = 2, 3, \ldots, n-1$, y $|a_{nn}| > |a_{n,n-1}|$, entonces A es no singular y los valores de l_{ii} descritos en el algoritmo de factorización Crout son diferentes a cero para cada $i = 1, 2, \ldots, n$. ■

La sección Conjunto de ejercicios 6.6 está disponible en línea. Encuentre la ruta de acceso en las páginas preliminares.

6.7 Software numérico

El software para operaciones de matriz y la solución directa de sistemas lineales implementados en IMSL y NAG está basado en LAPACK, un paquete de subrutinas de dominio público. Existe excelente documentación y libros disponible acerca de ellos. Nosotros nos enfocaremos en varias subrutinas disponibles en las tres fuentes.

LAPACK de acompañamiento es un conjunto de operaciones de nivel inferior llamadas Subprogramas Básicos de Álgebra Lineal (BLAS). En general, el nivel 1 de BLAS consiste en operaciones vector-vector, como sumas de vector con datos de entrada y conteo de operaciones de $O(n)$. El nivel 2 consiste en operaciones matriz-vector, como el producto de una matriz y un vector con datos de entrada y conteo de operaciones de $O(n^2)$. El nivel 3 consiste en operaciones matriz-matriz, como productos de matriz con datos de entrada y conteo de operaciones de $O(n^3)$.

Las subrutinas en LAPACK para resolver sistemas lineales primero factorizan la matriz A. La factorización depende del tipo de matriz de la siguiente forma:

1. Matriz general $PA = LU$

2. Matriz definida positiva $A = LL^t$

3. Matriz simétrica $A = LDL^t$

4. Matriz tridiagonal $A = LU$ (en forma de banda)

Además, se pueden calcular inversas y determinantes.

La Biblioteca IMSL incluye contrapartes para casi todas las subrutinas LAPACK y también algunas extensiones. La Biblioteca NAG tiene numerosas subrutinas para métodos directos para resolver sistemas lineales similares a los de LAPACK e IMSL.

Las secciones Preguntas de análisis, Conceptos clave y Revisión del capítulo están disponibles en línea. Encuentre la ruta de acceso en las páginas preliminares.

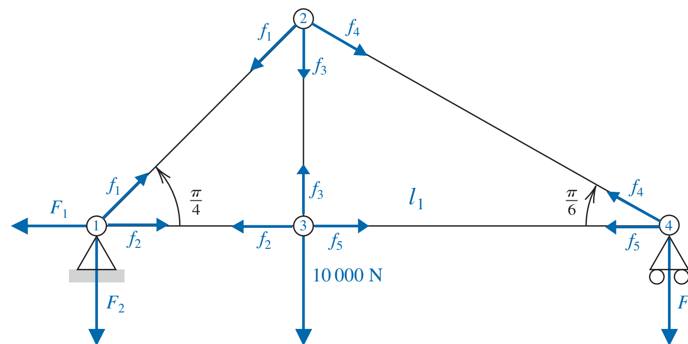
CAPÍTULO

7

Técnicas iterativas en álgebra de matrices

Introducción

Las vigas son estructuras livianas capaces de sostener cargas pesadas. En el diseño de puentes, las partes individuales de las vigas están conectadas con uniones orientables que permiten transferir las fuerzas desde una parte de la viga hasta otra. La figura adjunta muestra una viga sostenida de manera estacionaria en el extremo izquierdo inferior ① permite movimiento horizontal en el extremo derecho inferior ④ y tiene uniones en ①, ②, ③ y ④. Una carga de 10 000 newtons (N) se coloca en la unión ③ y las fuerzas resultantes en las uniones provistas por f_1, f_2, f_3, f_4 y f_5, como se muestra. Cuando son positivas, estas fuerzas indican tensión en los elementos de la viga y, cuando son negativas, compresión. El miembro de soporte estacionario podría tener tanto un componente de fuerza horizontal F_1 como un componente de fuerza vertical F_2, pero el miembro de soporte movible sólo tiene un componente de fuerza vertical F_3.



Si la viga está en equilibrio estático, las fuerzas en cada unión se deben sumar al vector cero, por lo que la suma de los componentes horizontales y verticales en cada unión debe ser 0. Esto produce el sistema lineal de ecuaciones que se muestra en la tabla adjunta. Una matriz de 8 × 8 que describe este sistema tiene 47 entradas cero y sólo 17 entradas diferentes a cero. Las matrices con un alto porcentaje de entradas cero reciben el nombre de *dispersas* y, a menudo, se resuelven utilizando métodos iterativos en lugar de técnicas directas. La solución iterativa a este sistema se considera en el ejercicio 15 de la sección 7.3 y en el ejercicio 10 en la sección 7.4.

Unión	Componente horizontal	Componente vertical
①	$-F_1 + \frac{\sqrt{2}}{2} f_1 + f_2 = 0$	$\frac{\sqrt{2}}{2} f_1 - F_2 = 0$
②	$-\frac{\sqrt{2}}{2} f_1 + \frac{\sqrt{3}}{2} f_4 = 0$	$-\frac{\sqrt{2}}{2} f_1 - f_3 - \frac{1}{2} f_4 = 0$
③	$-f_2 + f_5 = 0$	$f_3 - 10{,}000 = 0$
④	$-\frac{\sqrt{3}}{2} f_4 - f_5 = 0$	$\frac{1}{2} f_4 - F_3 = 0$

Los métodos que se presentan en el capítulo 6 usaban técnicas directas para resolver un sistema de $n \times n$ ecuaciones lineales de la forma $A\mathbf{x} = \mathbf{b}$. En este capítulo presentamos métodos iterativos para resolver un sistema de este tipo.



7.1 Normas de vectores y matrices

En el capítulo 2 describimos las técnicas iterativas para encontrar raíces de ecuaciones de la forma $f(x) = 0$. Se encontró una aproximación inicial (o aproximaciones) y, después, se determinaron aproximaciones nuevas con base en qué tan bien satisfacían la ecuación las aproximaciones previas. El objetivo es encontrar una forma de minimizar la diferencia entre las aproximaciones y la solución exacta.

Para analizar los métodos iterativos que resuelvan sistemas lineales, primero necesitamos determinar una forma para medir la distancia entre vectores columna n-dimensionales. Esto nos permitirá determinar si una sucesión de vectores converge a una solución del sistema.

Un escalar es un número real (o complejo) que por lo general se denota con letras itálicas o griegas. Los vectores se representan por medio de letras negritas.

En la actualidad, esta medida también se necesita cuando la solución se obtiene con los métodos directos presentados en el capítulo 6. Estos métodos requerían un gran número de operaciones aritméticas y utilizar aritmética de dígitos finitos sólo conduce a una aproximación para una solución real del sistema.

Normas de vector

Sea que $\mathbb{R}^n$, denota el conjunto de todos los vectores columna n-dimensionales con componentes de números reales. Para definir la distancia en $\mathbb{R}^n$, usamos la noción de una norma, que es la generalización del valor absoluto en $\mathbb{R}$, el conjunto de números reales.

Definición 7.1 Una **norma vectorial** en $\mathbb{R}^n$, es una función, $\|\cdot\|$, de $\mathbb{R}^n$, a $\mathbb{R}$, con las siguientes propiedades:

i) $\|\mathbf{x}\| \geq 0$ para toda $\mathbf{x} \in \mathbb{R}^n$,

ii) $\|\mathbf{x}\| = 0$ si y sólo si $\mathbf{x} = \mathbf{0}$,

iii) $\|\alpha\mathbf{x}\| = |\alpha|\|\mathbf{x}\|$ para toda $\alpha \in \mathbb{R}$ y $\mathbf{x} \in \mathbb{R}^n$,

iv) $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$ para toda $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$. ■

Los vectores en $\mathbb{R}^n$, son vectores columna y es conveniente usar la notación de la transpuesta presentada en la sección 6.3 cuando un vector se representa en términos de sus componentes. Por ejemplo, el vector

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

se escribirá $\mathbf{x} = (x_1, x_2, \ldots, x_n)^t$.

Sólo necesitaremos dos normas específicas en $\mathbb{R}^n$, a pesar de que se presenta una tercera norma en $\mathbb{R}^n$, en el ejercicio 9.

Definición 7.2 Las normas l_2 y l_∞ para el vector $\mathbf{x} = (x_1, x_2, \ldots, x_n)^t$ se definen mediante

$$\|\mathbf{x}\|_2 = \left\{\sum_{i=1}^{n} x_i^2\right\}^{1/2} \quad \text{y} \quad \|\mathbf{x}\|_\infty = \max_{1\leq i\leq n} |x_i|.$$ ■

Observe que cada una de estas normas se reduce al valor absoluto para el caso $n = 1$. La norma l_2 recibe el nombre de **norma euclidiana** del vector $\mathbf{x}$ porque representa la noción común de distancia desde el origen cuando $\mathbf{x}$ está en $\mathbb{R}^1 \equiv \mathbb{R}$, $\mathbb{R}^2$ o $\mathbb{R}^3$. Por ejemplo, la norma l_2 del vector $\mathbf{x} = (x_1, x_2, x_3)^t$ provee la longitud del segmento recto que une los puntos $(0, 0, 0)$ y (x_1, x_2, x_3). La figura 7.1 muestra la frontera de esos vectores en $\mathbb{R}^2$ y $\mathbb{R}^3$ que tienen una norma l_2 menor a 1. La figura 7.2 es una ilustración similar de la norma l_∞.

Figura 7.1

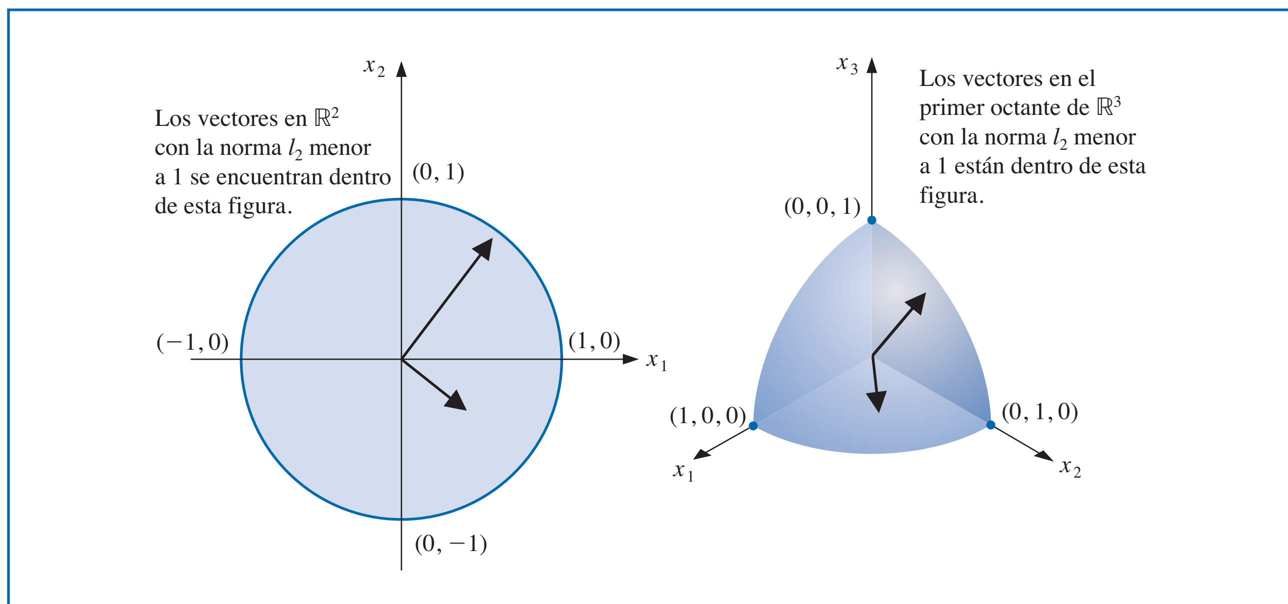
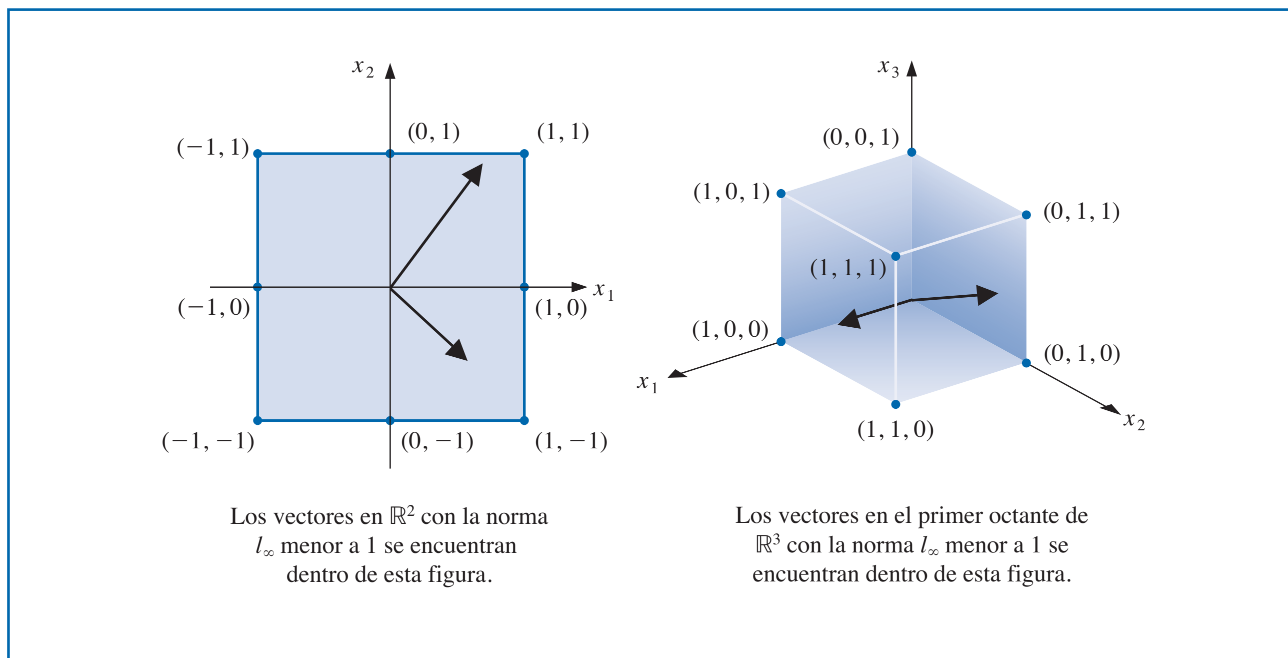


Figura 7.2



Ejemplo 1 Determine la norma l_2 y la norma l_∞ del vector $\mathbf{x} = (-1, 1, -2)^t$.

Solución El vector $\mathbf{x} = (-1, 1, -2)^t$ en $\mathbb{R}^3$ tiene las normas

$$\|\mathbf{x}\|_2 = \sqrt{(-1)^2 + (1)^2 + (-2)^2} = \sqrt{6}$$

y

$$\|\mathbf{x}\|_\infty = \text{máx}\{|-1|, |1|, |-2|\} = 2.$$

■

Es fácil mostrar que las propiedades en la definición 7.1 mantienen la norma l_∞ porque siguen resultados similares para valores absolutos. La única propiedad que requiere más demostración es iv) y, en este caso, si $\mathbf{x} = (x_1, x_2, \ldots, x_n)^t$ y $\mathbf{y} = (y_1, y_2, \ldots, y_n)^t$, entonces

$$\|\mathbf{x}+\mathbf{y}\|_\infty = \max_{1\le i\le n} |x_i + y_i| \le \max_{1\le i\le n} (|x_i| + |y_i|) \le \max_{1\le i\le n} |x_i| + \max_{1\le i\le n} |y_i| = \|\mathbf{x}\|_\infty + \|\mathbf{y}\|_\infty.$$

Las primeras tres condiciones también son fáciles de demostrar para la norma l_2. Pero para mostrar eso

$$\|\mathbf{x}+\mathbf{y}\|_2 \le \|\mathbf{x}\|_2 + \|\mathbf{y}\|_2, \quad \text{para cada } \mathbf{x}, \mathbf{y} \in \mathbb{R}_n,$$

necesitamos una desigualdad famosa.

Teorema 7.3 **(Desigualdad Cauchy-Bunyakovsky-Schwarz para sumas)**

Por cada $\mathbf{x} = (x_1, x_2, \ldots, x_n)^t$ y $\mathbf{y} = (y_1, y_2, \ldots, y_n)^t$ en $\mathbb{R}^n$,

$$\mathbf{x}^t\mathbf{y} = \sum_{i=1}^{n} x_i y_i \le \left\{\sum_{i=1}^{n} x_i^2\right\}^{1/2} \left\{\sum_{i=1}^{n} y_i^2\right\}^{1/2} = \|\mathbf{x}\|_2 \cdot \|\mathbf{y}\|_2. \tag{7.1}$$

Demostración Si $\mathbf{y} = \mathbf{0}$ o $\mathbf{x} = \mathbf{0}$, el resultado es inmediato porque ambos lados de la desigualdad son cero.

Existen muchas formas de esta desigualdad, por lo tanto, muchos descubridores. Augustin Louis Cauchy (1789–1857) la describió en 1821 en *Cours d'Analyse Algébrique (Curso de análisis algebraico)*, el primer libro sobre cálculo riguroso. Una integral de la forma de la igualdad aparece en el trabajo de Viktor Yakovlevich Bunyakovsky (1804–1889) en 1859 y Hermann Amandus Schwarz (1843–1921) usó una forma de integral doble de esta desigualdad en 1885. Más detalles sobre la historia se pueden encontrar en [Stee].

Suponga que $\mathbf{y} \ne \mathbf{0}$ y $\mathbf{x} \ne \mathbf{0}$. Observe que para cada $\lambda \in \mathbb{R}$ tenemos

$$0 \le ||\mathbf{x} - \lambda\mathbf{y}||_2^2 = \sum_{i=1}^{n} (x_i - \lambda y_i)^2 = \sum_{i=1}^{n} x_i^2 - 2\lambda \sum_{i=1}^{n} x_i y_i + \lambda^2 \sum_{i=1}^{n} y_i^2,$$

por lo que

$$2\lambda \sum_{i=1}^{n} x_i y_i \le \sum_{i=1}^{n} x_i^2 + \lambda^2 \sum_{i=1}^{n} y_i^2 = \|\mathbf{x}\|_2^2 + \lambda^2 \|\mathbf{y}\|_2^2.$$

Sin embargo $\|\mathbf{x}\|_2 > 0$ y $\|\mathbf{y}\|_2 > 0$, por lo que podemos hacer $\lambda = \|\mathbf{x}\|_2 / \|\mathbf{y}\|_2$ para obtener

$$\left(2\frac{\|\mathbf{x}\|_2}{\|\mathbf{y}\|_2}\right)\left(\sum_{i=1}^{n} x_i y_i\right) \le \|\mathbf{x}\|_2^2 + \frac{\|\mathbf{x}\|_2^2}{\|\mathbf{y}\|_2^2}\|\mathbf{y}\|_2^2 = 2\|\mathbf{x}\|_2^2.$$

Por lo tanto,

$$2\sum_{i=1}^{n} x_i y_i \le 2\|\mathbf{x}\|_2^2 \frac{\|\mathbf{y}\|_2}{\|\mathbf{x}\|_2} = 2\|\mathbf{x}\|_2\|\mathbf{y}\|_2,$$

y

$$\mathbf{x}^t\mathbf{y} = \sum_{i=1}^{n} x_i y_i \leq \|\mathbf{x}\|_2\|\mathbf{y}\|_2 = \left\{\sum_{i=1}^{n} x_i^2\right\}^{1/2}\left\{\sum_{i=1}^{n} y_i^2\right\}^{1/2}. \quad \blacksquare$$

Con este resultado vemos que para cada $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$,

$$\|\mathbf{x}+\mathbf{y}\|_2^2 = \sum_{i=1}^{n}(x_i + y_i)^2 = \sum_{i=1}^{n} x_i^2 + 2\sum_{i=1}^{n} x_i y_i + \sum_{i=1}^{n} y_i^2 \leq \|\mathbf{x}\|_2^2 + 2\|\mathbf{x}\|_2\|\mathbf{y}\|_2 + \|\mathbf{y}\|_2^2,$$

que proporciona la propiedad **iv)** de la norma:

$$\|\mathbf{x}+\mathbf{y}\|_2 \leq \left(\|\mathbf{x}\|_2^2 + 2\|\mathbf{x}\|_2\|\mathbf{y}\|_2 + \|\mathbf{y}\|_2^2\right)^{1/2} = \|\mathbf{x}\|_2 + \|\mathbf{y}\|_2.$$

Distancia entre vectores en $\mathbb{R}^n$

La norma de un vector proporciona una medida para la distancia entre un vector arbitrario y el vector cero, de la misma forma en la que el valor absoluto de un número real describe su distancia desde 0. De igual forma, la **distancia entre dos vectores** está definida como la norma de la diferencia de los vectores al igual que la distancia entre dos números reales es el valor absoluto de su diferencia.

Definición 7.4 Si $\mathbf{x} = (x_1, x_2, \ldots, x_n)^t$ y $\mathbf{y} = (y_1, y_2, \ldots, y_n)^t$ son vectores en $\mathbb{R}^n$, las distancias l_2 y l_∞ entre $\mathbf{x}$ y $\mathbf{y}$ se definen mediante

$$\|\mathbf{x}-\mathbf{y}\|_2 = \left\{\sum_{i=1}^{n}(x_i - y_i)^2\right\}^{1/2} \text{ y } \|\mathbf{x}-\mathbf{y}\|_\infty = \max_{1\leq i\leq n} |x_i - y_i|. \quad \blacksquare$$

Ejemplo 2 El sistema lineal

$$\begin{aligned}
3.3330x_1 + 15920x_2 - 10.333x_3 &= 15913,\\
2.2220x_1 + 16.710x_2 + 9.6120x_3 &= 28.544,\\
1.5611x_1 + 5.1791x_2 + 1.6852x_3 &= 8.4254,
\end{aligned}$$

tiene la solución exacta $\mathbf{x} = (x_1, x_2, x_3)^t = (1, 1, 1)^t$. La eliminación gaussiana realizada mediante aritmética de redondeo de cinco dígitos y pivoteo parcial (algoritmo 6.2) produce la solución aproximada

$$\tilde{\mathbf{x}} = (\tilde{x}_1, \tilde{x}_2, \tilde{x}_3)^t = (1.2001, 0.99991, 0.92538)^t.$$

Determine las distancias l_2 y l_∞ entre las soluciones exactas y aproximadas.

Solución Las mediciones de $\mathbf{x} - \tilde{\mathbf{x}}$ están dadas por

$$\begin{aligned}
\|\mathbf{x}-\tilde{\mathbf{x}}\|_\infty &= \text{máx}\{|1 - 1.2001|, |1 - 0.99991|, |1 - 0.92538|\}\\
&= \text{máx}\{0.2001, 0.00009, 0.07462\} = 0.2001
\end{aligned}$$

y

$$\begin{aligned}
\|\mathbf{x}-\tilde{\mathbf{x}}\|_2 &= \left[(1 - 1.2001)^2 + (1 - 0.99991)^2 + (1 - 0.92538)^2\right]^{1/2}\\
&= [(0.2001)^2 + (0.00009)^2 + (0.07462)^2]^{1/2} = 0.21356.
\end{aligned}$$

A pesar de que los componentes $\tilde{x}_2$ y $\tilde{x}_3$ son buenas aproximaciones para x_2 y x_3, el componente $\tilde{x}_1$ es una aproximación débil para x_1 y $|x_1 - \tilde{x}_1|$ domina ambas normas. ■

El concepto de distancia en $\mathbb{R}^n$ también se usa para definir una cota de una sucesión de vectores en este espacio.

Definición 7.5 Se dice que una sucesión $\{\mathbf{x}^{(k)}\}_{k=1}^{\infty}$ de vectores en $\mathbb{R}^n$ **converge** a $\mathbf{x}$ respecto a la norma $\|\cdot\|$ si, dado cualquier $\varepsilon > 0$, existe un entero $N(\varepsilon)$ tal que

$$\|\mathbf{x}^{(k)} - \mathbf{x}\| < \varepsilon, \quad \text{para toda } \ k \geq N(\varepsilon).$$ ■

Teorema 7.6 La sucesión de vectores $\{\mathbf{x}^{(k)}\}$ converge a $\mathbf{x}$ en $\mathbb{R}^n$ respecto a la norma l_∞ si y sólo si $\lim_{k\to\infty} x_i^{(k)} = x_i$, para cada $i = 1, 2, \ldots, n$.

Demostración Suponga que $\{\mathbf{x}^{(k)}\}$ converge a $\mathbf{x}$ respecto a la norma l_∞. Dado cualquier $\varepsilon > 0$, existe un entero $N(\varepsilon)$ tal que para toda $k \geq N(\varepsilon)$,

$$\max_{i=1,2,\ldots,n} |x_i^{(k)} - x_i| = \|\mathbf{x}^{(k)} - \mathbf{x}\|_\infty < \varepsilon.$$

Este resultado implica que $|x_i^{(k)} - x_i| < \varepsilon$, para cada $i = 1, 2, \ldots, n$, por lo que $\lim_{k\to\infty} x_i^{(k)} = x_i$ para cada i.

Por el contrario, suponga que $\lim_{k\to\infty} x_i^{(k)} = x_i$, para cada $i = 1, 2, \ldots, n$. Para un $\varepsilon > 0$, dado, sea $N_i(\varepsilon)$ para cada i representa un entero con la propiedad de que

$$|x_i^{(k)} - x_i| < \varepsilon,$$

siempre que $k \geq N_i(\varepsilon)$.

Defina $N(\varepsilon) = \max_{i=1,2,\ldots,n} N_i(\varepsilon)$. Si $k \geq N(\varepsilon)$, entonces

$$\max_{i=1,2,\ldots,n} |x_i^{(k)} - x_i| = \|\mathbf{x}^{(k)} - \mathbf{x}\|_\infty < \varepsilon.$$

Esto implica que $\{\mathbf{x}^{(k)}\}$ converge a $\mathbf{x}$ respecto a la norma l_∞. ■

Ejemplo 3 Muestre que

$$\mathbf{x}^{(k)} = (x_1^{(k)}, x_2^{(k)}, x_3^{(k)}, x_4^{(k)})^t = \left(1, 2 + \frac{1}{k}, \frac{3}{k^2}, e^{-k} \operatorname{sen} k\right)^t.$$

converge a $\mathbf{x} = (1, 2, 0, 0)^t$ respecto a la norma l_∞.

Solución Puesto que

$$\lim_{k\to\infty} 1 = 1, \quad \lim_{k\to\infty} (2 + 1/k) = 2, \quad \lim_{k\to\infty} 3/k^2 = 0 \quad \text{y} \quad \lim_{k\to\infty} e^{-k} \operatorname{sen} k = 0,$$

el teorema 7.6 implica que la sucesión $\{\mathbf{x}^{(k)}\}$ converge a $(1, 2, 0, 0)^t$ respecto a la norma l_∞. ■

Mostrar directamente que la sucesión en el ejemplo 3 converge a $(1, 2, 0, 0)^t$ respecto a la norma l_2 es bastante complicado. Es mejor probar el siguiente resultado y aplicarlo a este caso especial

Teorema 7.7 Para $\mathbf{x} \in \mathbb{R}^n$,

$$\|\mathbf{x}\|_\infty \leq \|\mathbf{x}\|_2 \leq \sqrt{n}\|\mathbf{x}\|_\infty.$$

Demostración Si x_j es una coordenada de $\mathbf{x}$ de tal forma que $\|\mathbf{x}\|_\infty = \text{máx}_{1\leq i\leq n} |x_i| = |x_j|$. Entonces

$$\|\mathbf{x}\|_\infty^2 = |x_j|^2 = x_j^2 \leq \sum_{i=1}^{n} x_i^2 = \|\mathbf{x}\|_2^2,$$

y

$$\|\mathbf{x}\|_\infty \leq \|\mathbf{x}\|_2.$$

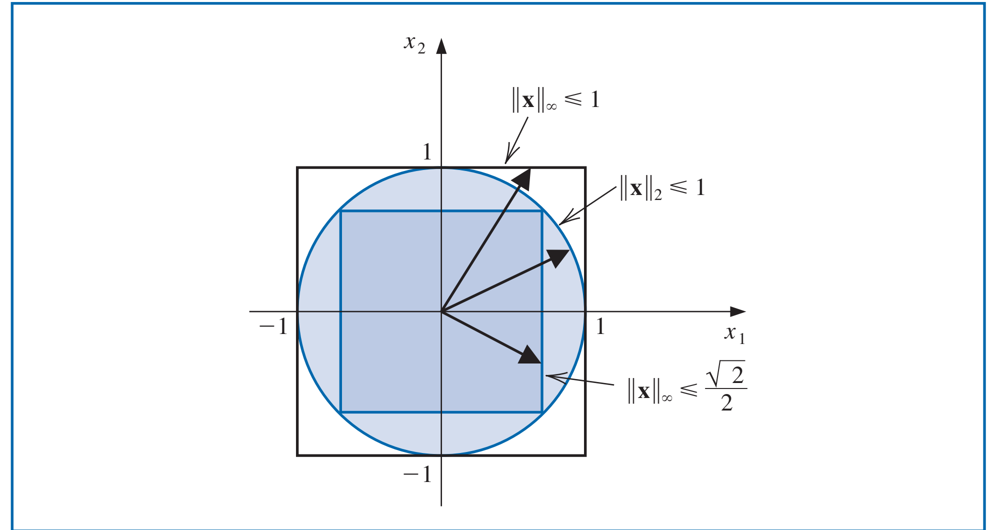
Por lo que,

$$\|\mathbf{x}\|_2^2 = \sum_{i=1}^{n} x_i^2 \leq \sum_{i=1}^{n} x_j^2 = nx_j^2 = n||\mathbf{x}||_\infty^2,$$

y $\|\mathbf{x}\|_2 \leq \sqrt{n}\|\mathbf{x}\|_\infty$. ■

La figura 7.3 ilustra este resultado cuando $n = 2$.

Figura 7.3



Ejemplo 4 En el ejemplo 3, encontramos que la sucesión $\{\mathbf{x}^{(k)}\}$, definida por

$$\mathbf{x}^{(k)} = \left(1, 2 + \frac{1}{k}, \frac{3}{k^2}, e^{-k} \operatorname{sen} k\right)^t,$$

converge a $\mathbf{x} = (1, 2, 0, 0)^t$ respecto a la norma l_∞. Muestre que esta sucesión también converge a $\mathbf{x}$ respecto a la norma l_2.

Solución Dado cualquier $\varepsilon > 0$, existe un entero $N(\varepsilon/2)$ con la propiedad de que

$$\|\mathbf{x}^{(k)} - \mathbf{x}\|_\infty < \frac{\varepsilon}{2},$$

siempre que $k \geq N(\varepsilon/2)$. Mediante el teorema 7.7, esto implica que

$$\|\mathbf{x}^{(k)} - \mathbf{x}\|_2 \leq \sqrt{4}\|\mathbf{x}^{(k)} - \mathbf{x}\|_\infty \leq 2(\varepsilon/2) = \varepsilon,$$

cuando $k \geq N(\varepsilon/2)$. Por lo que $\{\mathbf{x}^{(k)}\}$ también converge a $\mathbf{x}$ respecto a la norma l_2. ■

Se puede mostrar que todas las normas en $\mathbb{R}^n$ son equivalentes respecto a la convergencia; es decir, si $\|\cdot\|$ y $\|\cdot\|'$ son dos normas cualquiera sobre $\mathbb{R}^n$ y $\{\mathbf{x}^{(k)}\}_{k=1}^{\infty}$ tiene el límite $\mathbf{x}$ respecto a $\|\cdot\|$, entonces $\{\mathbf{x}^{(k)}\}_{k=1}^{\infty}$ también tiene el límite $\mathbf{x}$ respecto a $\|\cdot\|'$. La prueba de este hecho para el caso general se puede encontrar en [Or2], p. 8. El caso para las normas l_2 y l_∞ se siguen el teorema 7.7.

Normas matriciales y distancias

En las siguientes secciones de este capítulo y en capítulos posteriores, necesitaremos métodos para determinar la distancia entre matrices $n \times n$. Para ello se requiere el concepto de norma.

Definición 7.8 Una **norma matricial** sobre el conjunto de las matrices $n \times n$ es una función de valor real $\|\cdot\|$, definida en este conjunto, que se cumple para todas las matrices A y B $n \times n$ y todos los números reales α: ■

i) $\|A\| \geq 0$;

ii) $\|A\| = 0$, si y sólo si A es O, la matriz con todas las entradas 0;

iii) $\|\alpha A\| = |\alpha|\|A\|$;

iv) $\|A + B\| \leq \|A\| + \|B\|$;

v) $\|AB\| \leq \|A\|\|B\|$. ■

La **distancia entre matrices** A y B $\boldsymbol{n \times n}$ respecto a esta norma matricial es $\|A - B\|$.

A pesar de que las normas matriciales se pueden obtener de diversas maneras, las que se consideran con mayor frecuencia son las que son consecuencias naturales de las normas de vectores l_2 y l_∞.

Estas normas se definen al utilizar el siguiente teorema, cuya prueba se considera en el ejercicio 17.

Teorema 7.9 Si $\|\cdot\|$, es una norma vectorial en $\mathbb{R}^n$, entonces

$$\|A\| = \max_{\|\mathbf{x}\|=1} \|A\mathbf{x}\| \tag{7.2}$$

es una norma matricial. ■

Cada norma vectorial produce una norma matricial natural asociada.

Las normas matriciales definidas por normas vectoriales reciben el nombre de **normas matriciales naturales**, o *inducidas*, asociadas con la norma del vector. En este texto, se asumirá que todas las normas matriciales son naturales a menos que se especifique lo contrario.

Para cualquier $\mathbf{z} \neq \mathbf{0}$, el vector $\mathbf{x} = \mathbf{z}/\|\mathbf{z}\|$ es un vector unitario. Por lo tanto,

$$\max_{\|\mathbf{x}\|=1} \|A\mathbf{x}\| = \max_{\mathbf{z}\neq\mathbf{0}} \left\| A\left(\frac{\mathbf{z}}{\|\mathbf{z}\|}\right)\right\| = \max_{\mathbf{z}\neq\mathbf{0}} \frac{\|A\mathbf{z}\|}{\|\mathbf{z}\|},$$

y, alternativamente, podemos escribir

$$\|A\| = \max_{\mathbf{z}\neq\mathbf{0}} \frac{\|A\mathbf{z}\|}{\|\mathbf{z}\|}. \tag{7.3}$$

El siguiente corolario para el teorema 7.9 sigue esta representación de $\|A\|$.

Corolario 7.10 para cualquier vector $\mathbf{z} \neq \mathbf{0}$, matriz A y cualquier norma natural $\|\cdot\|$, tenemos

$$\|A\mathbf{z}\| \leq \|A\| \cdot \|\mathbf{z}\|.$$ ■

La medida dada a una matriz bajo la norma natural describe la forma en la que la matriz extiende los vectores unitarios relativos a esa norma. La extensión máxima es la norma de la matriz. Las normas de la matriz que consideraremos tienen las formas

$$\|A\|_\infty = \max_{\|\mathbf{x}\|_\infty = 1} \|A\mathbf{x}\|_\infty, \quad \text{la norma } l_\infty,$$

y

$$\|A\|_2 = \max_{\|\mathbf{x}\|_2 = 1} \|A\mathbf{x}\|_2, \quad \text{la norma } l_2.$$

Una ilustración de estas normas cuando $n = 2$ se muestra en las figuras 7.4 y 7.5 para la matriz

$$A = \begin{bmatrix} 0 & -2 \\ 2 & 0 \end{bmatrix}.$$

Figura 7.4

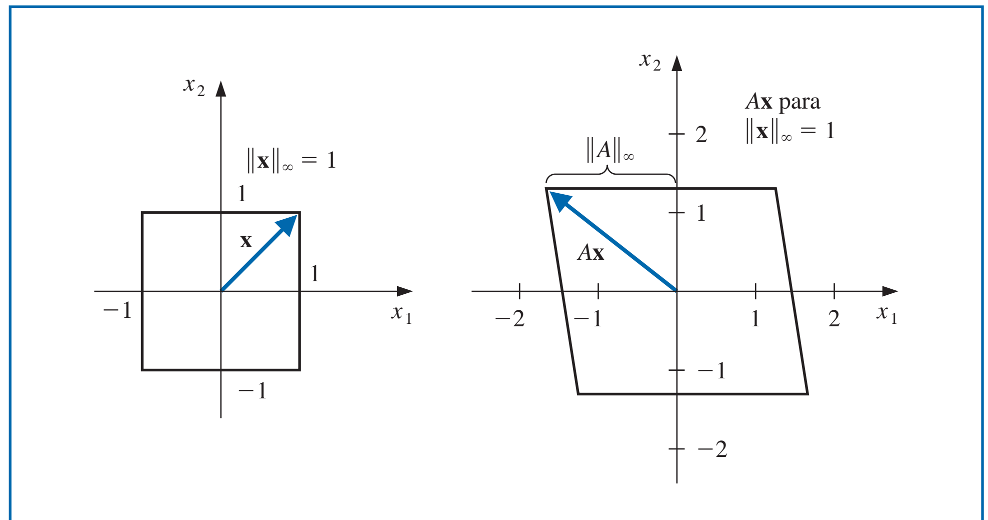
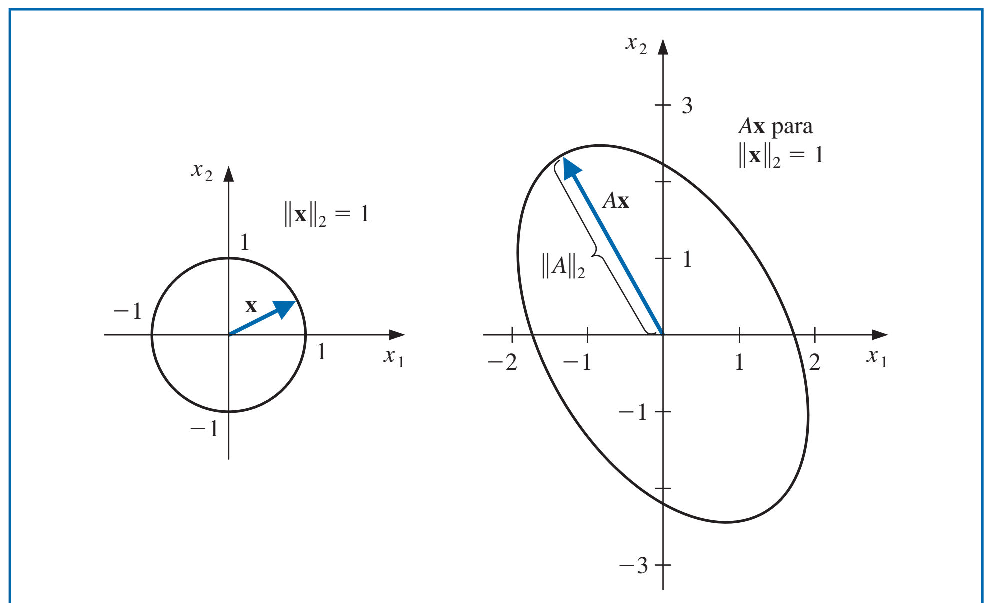


Figura 7.5



La norma l_∞ de una matriz se puede calcular fácilmente a partir de las entradas de la matriz.

Teorema 7.11 Si $A = (a_{ij})$ es una matriz $n \times n$, entonces $\|A\|_\infty = \max\limits_{1\le i\le n} \sum_{j=1}^{n} |a_{ij}|$.

Demostración Primero, mostramos que $\|A\|_\infty \le \max\limits_{1\le i\le n} \sum_{j=1}^{n} |a_{ij}|$.

Sea $\mathbf{x}$ un vector n-dimensional con $1 = \|\mathbf{x}\|_\infty = \text{máx}_{1\le i\le n} |x_i|$. Puesto que $A\mathbf{x}$ también es un vector n-dimensional

$$\|A\mathbf{x}\|_\infty = \max_{1\le i\le n} |(A\mathbf{x})_i| = \max_{1\le i\le n} \left| \sum_{j=1}^{n} a_{ij}x_j \right| \le \max_{1\le i\le n} \sum_{j=1}^{n} |a_{ij}| \max_{1\le j\le n} |x_j|.$$

Pero $\text{máx}_{1\le j\le n} |x_j| = \|\mathbf{x}\|_\infty = 1$, por lo que

$$\|A\mathbf{x}\|_\infty \le \max_{1\le i\le n} \sum_{j=1}^{n} |a_{ij}|,$$

y, por consiguiente,

$$\|A\|_\infty = \max_{\|\mathbf{x}\|_\infty=1} \|A\mathbf{x}\|_\infty \le \max_{1\le i\le n} \sum_{j=1}^{n} |a_{ij}|. \tag{7.4}$$

Ahora mostraremos la desigualdad opuesta. Si p es un entero con

$$\sum_{j=1}^{n} |a_{pj}| = \max_{1\le i\le n} \sum_{j=1}^{n} |a_{ij}|$$

y $\mathbf{x}$ es el vector con componentes

$$x_j = \begin{cases} 1, & \text{si } a_{pj} \ge 0, \\ -1, & \text{si } a_{pj} < 0. \end{cases}$$

Entonces $\|\mathbf{x}\|_\infty = 1$ y $a_{pj}x_j = |a_{pj}|$, para todas las $j = 1, 2, \ldots, n$, por lo que

$$\|A\mathbf{x}\|_\infty = \max_{1\le i\le n} \left| \sum_{j=1}^{n} a_{ij}x_j \right| \ge \left| \sum_{j=1}^{n} a_{pj}x_j \right| = \left| \sum_{j=1}^{n} |a_{pj}| \right| = \max_{1\le i\le n} \sum_{j=1}^{n} |a_{ij}|.$$

Este resultado implica que

$$\|A\|_\infty = \max_{\|\mathbf{x}\|_\infty=1} \|A\mathbf{x}\|_\infty \ge \max_{1\le i\le n} \sum_{j=1}^{n} |a_{ij}|.$$

Al unir esto con la desigualdad (7.4) obtenemos $\|A\|_\infty = \max\limits_{1\le i\le n} \sum_{j=1}^{n} |a_{ij}|$. ■

Ejemplo 5 Determine $\|A\|_\infty$ para la matriz

$$A = \begin{bmatrix} 1 & 2 & -1 \\ 0 & 3 & -1 \\ 5 & -1 & 1 \end{bmatrix}.$$

Solución Tenemos

$$\sum_{j=1}^{3} |a_{1j}| = |1| + |2| + |-1| = 4, \quad \sum_{j=1}^{3} |a_{2j}| = |0| + |3| + |-1| = 4,$$

y

$$\sum_{j=1}^{3} |a_{3j}| = |5| + |-1| + |1| = 7.$$

Por lo que el teorema 7.11 implica que $\|A\|_\infty = \text{máx}\{4, 4, 7\} = 7$. ■

En la siguiente sección, veremos un método alternativo para encontrar la norma l_2 de una matriz.

La sección Conjunto de ejercicios 7.1 está disponible en línea. Encuentre la ruta de acceso en las páginas preliminares.

7.2 Eigenvalores y eigenvectores

Una matriz $n \times m$ se puede considerar como una función que utiliza multiplicación de matrices para tomar vectores columna m-dimensionales en vectores columna n-dimensionales. Por lo que, una matriz $n \times m$ es, en realidad, una función lineal de $\mathbb{R}^m$ a $\mathbb{R}^n$. Una matriz cuadrada A toma el conjunto de vectores n-dimensionales en sí misma, lo cual provee una función lineal de $\mathbb{R}^m$ a $\mathbb{R}^n$. En este caso, ciertos vectores $\mathbf{x}$ diferentes de cero podrían ser paralelos a $A\mathbf{x}$, lo que significa que existe una constante λ con $A\mathbf{x} = \lambda\mathbf{x}$. Para estos vectores, tenemos $(A - \lambda I)\mathbf{x} = \mathbf{0}$. Existe una conexión cercana entre los valores de λ y la probabilidad de que un método iterativo convergerá. Nosotros consideraremos la conexión en esta sección.

Definición 7.12 Si A es una matriz cuadrada, el **polinomio característico** de A está definido por

$$p(\lambda) = \det(A - \lambda I).$$ ■

No es difícil demostrar (consulte el ejercicio 15) que p es un polinomio de enésimo grado y, por consiguiente, tiene por lo menos n ceros diferentes, algunos de los cuales podrían ser complejos. Si λ es un cero de p, entonces, puesto que $\det(A - \lambda I) = 0$, el teorema 6.17 en la página 298 implica que el sistema lineal definido por $(A - \lambda I)\mathbf{x} = \mathbf{0}$ tiene una solución con $\mathbf{x} \neq \mathbf{0}$. Nos gustaría estudiar los ceros de p y las soluciones diferentes a cero correspondientes a estos sistemas.

Definición 7.13 Si p es el polinomio característico de la matriz A, los ceros de p reciben el nombre de **eigenvalores**, o valores característicos, de la matriz A. Si λ es un eigenvalor de A y $\mathbf{x} \neq \mathbf{0}$ satisface $(A - \lambda I)\mathbf{x} = \mathbf{0}$, entonces $\mathbf{x}$ es un **eigenvector**, o vector característico, de A correspondiente al eigenvalor λ. ■

El prefijo *eigen* proviene del adjetivo alemán que significa "propio" y en inglés es sinónimo de la palabra *característico*. Cada matriz tiene una eigenecuación o característica propia, con eigenvalores o funciones característicos correspondientes.

Para determinar los eigenvalores de una matriz, podemos utilizar el hecho de que

- λ es un eigenvalor de A si y sólo si $\det(A - \lambda I) = 0$.

Una vez que se ha encontrado el eigenvalor λ, un eigenvector correspondiente $\mathbf{x} \neq \mathbf{0}$ se determina al resolver el sistema

- $(A - \lambda I)\mathbf{x} = \mathbf{0}$.

Ejemplo 1 Muestre que no hay vectores $\mathbf{x}$ diferentes a cero en $\mathbb{R}^2$ con $A\mathbf{x}$ paralelo a $\mathbf{x}$ si

$$A = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}.$$

Solución Los eigenvalores de A son los ceros del polinomio característico

$$0 = \det(A - \lambda I) = \det \begin{bmatrix} -\lambda & 1 \\ -1 & -\lambda \end{bmatrix} = \lambda^2 + 1,$$

por lo que los eigenvalores de A son los números complejos $\lambda_1 = i$ y $\lambda_2 = -i$. Un eigenvector $\mathbf{x}$ correspondiente a λ_1 necesita satisfacer

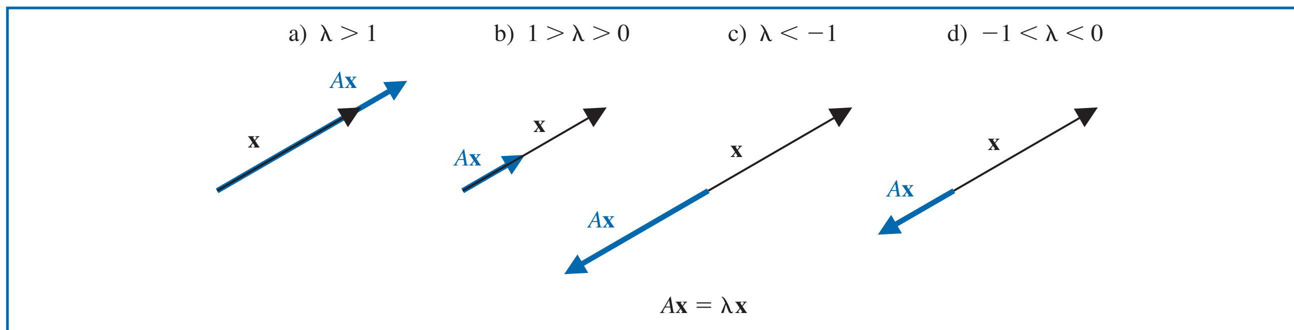
$$\begin{bmatrix} 0 \\ 0 \end{bmatrix} = \begin{bmatrix} -i & 1 \\ -1 & -i \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} -ix_1 + x_2 \\ -x_1 - ix_2 \end{bmatrix},$$

es decir, $0 = -ix_1 + x_2$, por lo que $x_2 = ix_1$, y $0 = -x_1 - ix_2$. Por lo tanto, si $\mathbf{x}$ es un eigenvector de A, entonces exactamente uno de sus componentes es real y el otro es complejo. Por consiguiente, no hay vectores $\mathbf{x}$ diferentes de cero en $\mathbb{R}^2$ con $A\mathbf{x}$ paralela a $\mathbf{x}$. ■

Si $\mathbf{x}$ es un eigenvector asociado con el eigenvalor real λ, entonces $A\mathbf{x} = \lambda\mathbf{x}$, por lo que la matriz A transforma el vector $\mathbf{x}$ en un múltiplo escalar de sí mismo.

- Si λ es real y $\lambda > 1$, entonces A tiene el efecto de expandir $\mathbf{x}$ en un factor de λ, como se ilustra en la figura 7.6 a).
- Si $0 < \lambda < 1$, entonces A comprime $\mathbf{x}$ en un factor de λ (consulte la figura 7.6 b)).
- Si $\lambda < 0$, los efectos son similares (consulte la figura 7.6 c) y d)), a pesar de que la dirección de $A\mathbf{x}$ está invertida.

Figura 7.6



También observe que si $\mathbf{x}$ es un eigenvector de A asociado con el eigenvalor λ y α es cualquier constante diferente a cero, entonces $\alpha\mathbf{x}$ también es un eigenvector ya que

$$A(\alpha\mathbf{x}) = \alpha(A\mathbf{x}) = \alpha(\lambda\mathbf{x}) = \lambda(\alpha\mathbf{x}).$$

Una consecuencia importante de esto es que para cualquier norma vectorial $||\cdot||$, podríamos seleccionar $\alpha = \pm||\mathbf{x}||^{-1}$, lo cual resultaría en $\alpha\mathbf{x}$ como el eigenvector con norma 1. Por lo que,

- Para todos los eigenvalores y cualquier norma vectorial, existen eigenvectores con norma 1.

Ejemplo 2 Determine los eigenvalores y eigenvectores para la matriz

$$A = \begin{bmatrix} 2 & 0 & 0 \\ 1 & 1 & 2 \\ 1 & -1 & 4 \end{bmatrix}.$$

Solución El polinomio característico de A es

$$p(\lambda) = \det(A - \lambda I) = \det \begin{bmatrix} 2-\lambda & 0 & 0 \\ 1 & 1-\lambda & 2 \\ 1 & -1 & 4-\lambda \end{bmatrix}$$

$$= -(\lambda^3 - 7\lambda^2 + 16\lambda - 12) = -(\lambda - 3)(\lambda - 2)^2,$$

por lo que existen dos eigenvalores de A: $\lambda_1 = 3$ y $\lambda_2 = 2$.

Un eigenvector $\mathbf{x}_1$ correspondiente al eigenvalor $\lambda_1 = 3$ es una solución para la ecuación de vector-matriz $(A - 3 \cdot I)\mathbf{x}_1 = \mathbf{0}$, por lo que

$$\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} -1 & 0 & 0 \\ 1 & -2 & 2 \\ 1 & -1 & 1 \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix},$$

lo cual implica que $x_1 = 0$ y $x_2 = x_3$.

Cualquier valor diferente de cero de x_3 produce un eigenvector para el eigenvalor $\lambda_1 = 3$. Por ejemplo, cuando $x_3 = 1$, tenemos el eigenvector $\mathbf{x}_1 = (0, 1, 1)^t$, y cualquier eigenvector de A correspondiente a $\lambda = 3$ es un múltiplo diferente a cero de $\mathbf{x}_1$.

Un eigenvector $\mathbf{x} \neq \mathbf{0}$ de A asociado con $\lambda_2 = 2$ es una solución del sistema $(A - 2 \cdot I)\mathbf{x} = \mathbf{0}$, por lo que

$$\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 \\ 1 & -1 & 2 \\ 1 & -1 & 2 \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}.$$

En este caso, el eigenvector sólo tiene que satisfacer la ecuación

$$x_1 - x_2 + 2x_3 = 0,$$

lo cual se puede realizar de diferentes formas. Por ejemplo, cuando $x_1 = 0$, tenemos $x_2 = 2x_3$, por lo que una elección sería $\mathbf{x}_2 = (0, 2, 1)^t$. También podríamos seleccionar $x_2 = 0$, lo cual requiere que $x_1 = -2x_3$. Por lo tanto, $\mathbf{x}_3 = (-2, 0, 1)^t$ da un segundo eigenvector para el eigenvalor $\lambda_2 = 2$ que no es un múltiplo de $\mathbf{x}_2$. Los eigenvectores de A correspondientes al eigenvalor $\lambda_2 = 2$ generan un plano entero. Este plano se describe mediante todos los vectores de la forma

$$\alpha\mathbf{x}_2 + \beta\mathbf{x}_3 = (-2\beta, 2\alpha, \alpha + \beta)^t,$$

para constantes arbitrarias α y β, siempre y cuando al menos una de las constantes sea diferente de cero. ■

Las nociones de los eigenvalores y los eigenvectores se introducen aquí para conveniencia computacional específica, pero estos conceptos surgen con frecuencia en el estudio de sistemas físicos. De hecho, puesto que son bastante interesantes el capítulo 9 está dedicado a su aproximación numérica.

Radio espectral

Definición 7.14 El **radio espectral** $\rho(A)$ de una matriz A está definido por

$$\rho(A) = \text{máx}\,|\lambda|, \text{ donde } \lambda \text{ es un eigenvalor de } A.$$

(Para $\lambda = \alpha + \beta i$, complejo, definimos $|\lambda| = (\alpha^2 + \beta^2)^{1/2}$.) ■

Para la matriz considerada en el ejemplo 2, $\rho(A) = \text{máx}\{2, 3\} = 3$.

El radio espectral está estrechamente relacionado con la norma de una matriz, como se muestra en el siguiente teorema.

Teorema 7.15 Si A es una matriz $n \times n$, entonces

i) $\|A\|_2 = [\rho(A^t A)]^{1/2}$,

ii) $\rho(A) \leq \|A\|$, para cualquier norma natural $\|\cdot\|$.

Demostración La demostración de la parte i) requiere más información respecto a los eigenvalores actualmente disponibles. Para los detalles relacionados con la demostración, consulte [Or2], p. 21.

Para probar la parte ii), suponga que λ es un eigenvalor de A con eigenvector $\mathbf{x}$ y $\|\mathbf{x}\| = 1$. Entonces $A\mathbf{x} = \lambda\mathbf{x}$ y

$$|\lambda| = |\lambda| \cdot \|\mathbf{x}\| = \|\lambda\mathbf{x}\| = \|A\mathbf{x}\| \leq \|A\|\|\mathbf{x}\| = \|A\|.$$

Por lo tanto,

$$\rho(A) = \text{máx}\, |\lambda| \leq \|A\|.$$ ■

La parte i) del teorema 7.15 implica que si A es simétrica, entonces $\|A\|_2 = \rho(A)$ (consulte el ejercicio 18.)

Un resultado interesante y útil, que es similar a la parte ii) del teorema 7.15, es que para cualquier matriz A y cualquier $\varepsilon > 0$, existe una norma natural $\|\cdot\|$ con la propiedad de que $\rho(A) < \|A\| < \rho(A) + \varepsilon$. Por consiguiente, $\rho(A)$ es la cota inferior más grande para las normas naturales en A. La prueba de este resultado se puede encontrar en [Or2], p. 23.

Ejemplo 3 Determine la norma l_2 de

$$A = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 2 & 1 \\ -1 & 1 & 2 \end{bmatrix}.$$

Solución Para aplicar el teorema 7.15, necesitamos calcular $\rho(A^t A)$, por lo que primero necesitamos los eigenvalores de $A^t A$:

$$A^t A = \begin{bmatrix} 1 & 1 & -1 \\ 1 & 2 & 1 \\ 0 & 1 & 2 \end{bmatrix} \begin{bmatrix} 1 & 1 & 0 \\ 1 & 2 & 1 \\ -1 & 1 & 2 \end{bmatrix} = \begin{bmatrix} 3 & 2 & -1 \\ 2 & 6 & 4 \\ -1 & 4 & 5 \end{bmatrix}.$$

Si

$$0 = \det(A^t A - \lambda I) = \det \begin{bmatrix} 3-\lambda & 2 & -1 \\ 2 & 6-\lambda & 4 \\ -1 & 4 & 5-\lambda \end{bmatrix}$$

$$= -\lambda^3 + 14\lambda^2 - 42\lambda = -\lambda(\lambda^2 - 14\lambda + 42),$$

mediante $\lambda = 0$ o $\lambda = 7 \pm \sqrt{7}$. Mediante el teorema 7.15, tenemos

$$||A||_2 = \sqrt{\rho(A^t A)} = \sqrt{\text{máx}\{0, 7-\sqrt{7}, 7+\sqrt{7}\}} = \sqrt{7+\sqrt{7}} \approx 3.106.$$ ■

Matrices convergentes

Al estudiar técnicas de matrices iterativas, es especialmente importante saber cuándo las potencias de una matriz se vuelven pequeñas (es decir, cuando todas las entradas se aproximan a cero). Las matrices de este tipo reciben el nombre de *convergentes*.

Definición 7.16 Llamamos **convergente** a una matriz A $n \times n$ si

$$\lim_{k\to\infty} (A^k)_{ij} = 0, \quad \text{para cada } i = 1, 2, \ldots, n \text{ y } j = 1, 2, \ldots, n. \qquad \blacksquare$$

Ejemplo 4 Muestre que

$$A = \begin{bmatrix} \frac{1}{2} & 0 \\ \frac{1}{4} & \frac{1}{2} \end{bmatrix}$$

es una matriz convergente.

Solución Al calcular las potencias de A, obtenemos

$$A^2 = \begin{bmatrix} \frac{1}{4} & 0 \\ \frac{1}{4} & \frac{1}{4} \end{bmatrix}, \quad A^3 = \begin{bmatrix} \frac{1}{8} & 0 \\ \frac{3}{16} & \frac{1}{8} \end{bmatrix}, \quad A^4 = \begin{bmatrix} \frac{1}{16} & 0 \\ \frac{1}{8} & \frac{1}{16} \end{bmatrix},$$

y, en general,

$$A^k = \begin{bmatrix} (\frac{1}{2})^k & 0 \\ \frac{k}{2^{k+1}} & (\frac{1}{2})^k \end{bmatrix}.$$

Por lo que A es una matriz convergente porque

$$\lim_{k\to\infty} \left(\frac{1}{2}\right)^k = 0 \quad \text{y} \quad \lim_{k\to\infty} \frac{k}{2^{k+1}} = 0. \qquad \blacksquare$$

Observe que la matriz convergente A en el ejemplo 4 tiene $\rho(A) = \frac{1}{2}$ porque $\frac{1}{2}$ es el único eigenvalor de A. Esto ilustra una conexión importante que existe entre el radio espectral de una matriz y la convergencia de la matriz, como se detalla en el siguiente resultado.

Teorema 7.17 Las siguientes declaraciones son equivalentes

i) A es una matriz convergente

ii) $\lim_{n\to\infty} \|A^n\| = 0$, para alguna norma natural.

iii) $\lim_{n\to\infty} \|A^n\| = 0$, para todas las normas naturales.

iv) $\rho(A) < 1$.

v) $\lim_{n\to\infty} A^n\mathbf{x} = \mathbf{0}$, para cada $\mathbf{x}$. $\blacksquare$

La demostración de este teorema se puede encontrar en [IK], p. 14.

La sección Conjunto de ejercicios 7.2 está disponible en línea. Encuentre la ruta de acceso en las páginas preliminares.

7.3 Técnicas iterativas de Jacobi y Gauss-Siedel

En esta sección describimos los métodos iterativos de Jacobi y Gauss-Siedel, métodos clásicos que datan de finales del siglo XVIII. Las técnicas iterativas casi nunca se usan para resolver sistemas lineales de dimensiones pequeñas ya que el tiempo requerido para suficiente precisión excede el requerido para las técnicas directas, como la eliminación gaussiana. Para grandes sistemas con un alto porcentaje de entradas 0, sin embargo, estas técnicas son eficientes en términos tanto de almacenamiento como de cálculo computacional. Los sistemas de este tipo surgen con frecuencia en análisis de circuitos y en la solución numérica de problemas de valor en la frontera y ecuaciones diferenciales parciales.

Una técnica iterativa para resolver el sistema lineal $n \times n$ $A\mathbf{x} = \mathbf{b}$ inicia con una aproximación $\mathbf{x}^{(0)}$ para la solución $\mathbf{x}$ y genera una sucesión de vectores $\{\mathbf{x}^{(k)}\}_{k=0}^{\infty}$ que convergen a $\mathbf{x}$.

Método de Jacobi

El **método iterativo de Jacobi** se obtiene al resolver la i-ésima ecuación en $A\mathbf{x} = \mathbf{b}$ para x_i para obtener (siempre que $a_{ii} \neq 0$)

$$x_i = \sum_{\substack{j=1 \\ j\neq i}}^{n} \left(-\frac{a_{ij}x_j}{a_{ii}} \right) + \frac{b_i}{a_{ii}}, \qquad \text{para } i = 1, 2, \ldots, n.$$

Para cada $k \geq 1$, genere los componentes $x_i^{(k)}$ de $\mathbf{x}^{(k)}$ a partir de los componentes de $\mathbf{x}^{(k-1)}$ por medio de

$$x_i^{(k)} = \frac{1}{a_{ii}} \left[\sum_{\substack{j=1 \\ j\neq i}}^{n} \left(-a_{ij}x_j^{(k-1)} \right) + b_i \right], \qquad \text{para } i = 1, 2, \ldots, n. \tag{7.5}$$

Ejemplo 1 El sistema lineal $A\mathbf{x} = \mathbf{b}$ dado por

$$\begin{aligned}
E_1: &\quad 10x_1 - x_2 + 2x_3 = 6,\\
E_2: &\quad -x_1 + 11x_2 - x_3 + 3x_4 = 25,\\
E_3: &\quad 2x_1 - x_2 + 10x_3 - x_4 = -11,\\
E_4: &\quad 3x_2 - x_3 + 8x_4 = 15
\end{aligned}$$

Carl Gustav Jacob Jacobi (1804–1851) fue reconocido en primer lugar por su trabajo en el área de la teoría de números y funciones elípticas, pero sus intereses matemáticos y capacidades eran muy amplias. Tenía una personalidad fuerte que influyó en el establecimiento de una actitud orientada hacia la investigación que se convirtió en el centro del resurgimiento de las matemáticas en las universidades alemanas en el siglo XIX.

tiene la solución única $\mathbf{x} = (1, 2, -1, 1)^t$. Utilice la técnica iterativa de Jacobi para encontrar aproximaciones $\mathbf{x}^{(k)}$ para $\mathbf{x}$, que inicia con $\mathbf{x}^{(0)} = (0, 0, 0, 0)^t$ hasta

$$\frac{\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\|_\infty}{\|\mathbf{x}^{(k)}\|_\infty} < 10^{-3}.$$

Solución Primero resolvemos la ecuación E_i para x_i, para cada $i = 1, 2, 3, 4$, a fin de obtener

$$\begin{aligned}
x_1 &= \frac{1}{10}x_2 - \frac{1}{5}x_3 + \frac{3}{5},\\
x_2 &= \frac{1}{11}x_1 + \frac{1}{11}x_3 - \frac{3}{11}x_4 + \frac{25}{11},\\
x_3 &= -\frac{1}{5}x_1 + \frac{1}{10}x_2 + \frac{1}{10}x_4 - \frac{11}{10},\\
x_4 &= -\frac{3}{8}x_2 + \frac{1}{8}x_3 + \frac{15}{8}.
\end{aligned}$$

A partir de la aproximación inicial $\mathbf{x}^{(0)} = (0, 0, 0, 0)^t$ tenemos $\mathbf{x}^{(1)}$ provista por

$$\begin{aligned}
x_1^{(1)} &= \quad\quad\quad\quad\;\; \frac{1}{10}x_2^{(0)} - \frac{1}{5}x_3^{(0)} \quad\quad\quad\; + \frac{3}{5} = \;\;0.6000,\\
x_2^{(1)} &= \frac{1}{11}x_1^{(0)} \quad\quad\quad\;\; + \frac{1}{11}x_3^{(0)} - \frac{3}{11}x_4^{(0)} + \frac{25}{11} = \;\;2.2727,\\
x_3^{(1)} &= -\frac{1}{5}x_1^{(0)} + \frac{1}{10}x_2^{(0)} \quad\quad\quad\; + \frac{1}{10}x_4^{(0)} - \frac{11}{10} = -1.1000,\\
x_4^{(1)} &= \quad\quad\quad\; - \frac{3}{8}x_2^{(0)} + \frac{1}{8}x_3^{(0)} \quad\quad\quad\; + \frac{15}{8} = \;\;1.8750.
\end{aligned}$$

Las iteraciones adicionales, $\mathbf{x}^{(k)} = (x_1^{(k)}, x_2^{(k)}, x_3^{(k)}, x_4^{(k)})^t$, se generan de manera similar y se presentan en la tabla 7.1.

Tabla 7.1

k	0	1	2	3	4	5	6	7	8	9	10
$x_1^{(k)}$	0.000	0.6000	1.0473	0.9326	1.0152	0.9890	1.0032	0.9981	1.0006	0.9997	1.0001
$x_2^{(k)}$	0.0000	2.2727	1.7159	2.053	1.9537	2.0114	1.9922	2.0023	1.9987	2.0004	1.9998
$x_3^{(k)}$	0.0000	−1.1000	−0.8052	−1.0493	−0.9681	−1.0103	−0.9945	−1.0020	−0.9990	−1.0004	−0.9998
$x_4^{(k)}$	0.0000	1.8750	0.8852	1.1309	0.9739	1.0214	0.9944	1.0036	0.9989	1.0006	0.9998

Nos detuvimos después de 10 iteraciones porque

$$\frac{\|\mathbf{x}^{(10)} - \mathbf{x}^{(9)}\|_\infty}{\|\mathbf{x}^{(10)}\|_\infty} = \frac{8.0 \times 10^{-4}}{1.9998} < 10^{-3}.$$

De hecho, $\|\mathbf{x}^{(10)} - \mathbf{x}\|_\infty = 0.0002$. ■

En general, las técnicas iterativas para resolver sistemas lineales implican un proceso que convierte el sistema $A\mathbf{x} = \mathbf{b}$ en un sistema equivalente de la forma $\mathbf{x} = T\mathbf{x} + \mathbf{c}$ para una matriz fija T y vector $\mathbf{c}$. Después de seleccionar el vector inicial $\mathbf{x}^{(0)}$, la sucesión de los vectores solución aproximados se generan al calcular

$$\mathbf{x}^{(k)} = T\mathbf{x}^{(k-1)} + \mathbf{c},$$

para cada $k = 1, 2, 3, \ldots$ Esto debería recordar la iteración de punto fijo estudiada en el capítulo 2.

El método de Jacobi se puede escribir en la forma $\mathbf{x}^{(k)} = T\mathbf{x}^{(k-1)} + \mathbf{c}$ al dividir A en sus partes diagonal o fuera de la diagonal. Para observar esto, permita que D sea la matriz diagonal cuyas entradas diagonales sean las de A, $-L$ es la parte estrictamente triangular inferior de A y $-U$ es la parte estrictamente triangular superior de A. Con esta notación,

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix}$$

se divide en

$$A = \begin{bmatrix} a_{11} & 0 & \cdots & 0 \\ 0 & a_{22} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & a_{nn} \end{bmatrix} - \begin{bmatrix} 0 & \cdots & \cdots & 0 \\ -a_{21} & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ -a_{n1} & \cdots & -a_{n,n-1} & 0 \end{bmatrix} - \begin{bmatrix} 0 & -a_{12} & \cdots & -a_{1n} \\ \vdots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & -a_{n-1,n} \\ 0 & \cdots & \cdots & 0 \end{bmatrix}$$

$$= D - L - U.$$

Entonces, la ecuación $A\mathbf{x} = \mathbf{b}$, o $(D - L - U)\mathbf{x} = \mathbf{b}$, se transforma en

$$D\mathbf{x} = (L + U)\mathbf{x} + \mathbf{b},$$

y, si existe D^{-1}, es decir, si $a_{ii} \neq 0$ para cada i, entonces

$$\mathbf{x} = D^{-1}(L + U)\mathbf{x} + D^{-1}\mathbf{b}.$$

Esto resulta en forma matricial de la técnica iterativa de Jacobi:

$$\mathbf{x}^{(k)} = D^{-1}(L + U)\mathbf{x}^{(k-1)} + D^{-1}\mathbf{b}, \quad k = 1, 2, \ldots. \tag{7.6}$$

Al introducir la notación $T_j = D^{-1}(L + U)$ y $\mathbf{c}_j = D^{-1}\mathbf{b}$ obtenemos la forma de la técnica de Jacobi

$$\mathbf{x}^{(k)} = T_j\mathbf{x}^{(k-1)} + \mathbf{c}_j. \tag{7.7}$$

En la práctica, la ecuación (7.5) se usa en el cálculo y la ecuación (7.7) para propósitos teóricos.

Ejemplo 2 Exprese el método de iteración de Jacobi para el sistema lineal $A\mathbf{x} = \mathbf{b}$ dado por

$$\begin{aligned} E_1: &\quad 10x_1 - x_2 + 2x_3 &= 6, \\ E_2: &\quad -x_1 + 11x_2 - x_3 + 3x_4 &= 25, \\ E_3: &\quad 2x_1 - x_2 + 10x_3 - x_4 &= -11, \\ E_4: &\quad 3x_2 - x_3 + 8x_4 &= 15, \end{aligned}$$

en la forma $\mathbf{x}^{(k)} = T\mathbf{x}^{(k-1)} + \mathbf{c}$.

Solución En el ejemplo 1 observamos que el método de Jacobi para este sistema tiene la forma

$$\begin{aligned} x_1 &= \frac{1}{10}x_2 - \frac{1}{5}x_3 + \frac{3}{5}, \\ x_2 &= \frac{1}{11}x_1 + \frac{1}{11}x_3 - \frac{3}{11}x_4 + \frac{25}{11}, \\ x_3 &= -\frac{1}{5}x_1 + \frac{1}{10}x_2 + \frac{1}{10}x_4 - \frac{11}{10}, \\ x_4 &= -\frac{3}{8}x_2 + \frac{1}{8}x_3 + \frac{15}{8}. \end{aligned}$$

Por lo tanto, tenemos

$$T = \begin{bmatrix} 0 & \frac{1}{10} & -\frac{1}{5} & 0 \\ \frac{1}{11} & 0 & \frac{1}{11} & -\frac{3}{11} \\ -\frac{1}{5} & \frac{1}{10} & 0 & \frac{1}{10} \\ 0 & -\frac{3}{8} & \frac{1}{8} & 0 \end{bmatrix} \quad \text{y} \quad \mathbf{c} = \begin{bmatrix} \frac{3}{5} \\ \frac{25}{11} \\ -\frac{11}{10} \\ \frac{15}{8} \end{bmatrix}.$$ ■

El algoritmo 7.1 implementa la técnica iterativa de Jacobi.

ALGORITMO
7.1

Técnica iterativa de Jacobi

Para resolver $A\mathbf{x} = \mathbf{b}$ dada una aproximación inicial $\mathbf{x}^{(0)}$:

ENTRADA el número de ecuaciones y valores desconocidos n; las entradas a_{ij}, $1 \leq i, j \leq n$ de la matriz A; las entradas b_i, $1 \leq i \leq n$ de $\mathbf{b}$; las entradas XO_i, $1 \leq i \leq n$ de $\mathbf{XO} = \mathbf{x}^{(0)}$; tolerancia TOL; número máximo de iteraciones N.

SALIDA la solución aproximada $x_1, \ldots, x_n$ o un mensaje que indica que se excedió el número de iteraciones.

Paso 1 Determine $k = 1$.

Paso 2 Mientras ($k \leq N$) haga los pasos 3–6.

Paso 3 Para $i = 1, \ldots, n$

$$\text{determine } x_i = \frac{1}{a_{ii}}\left[-\sum_{\substack{j=1\\ j\neq i}}^{n} (a_{ij}XO_j) + b_i\right].$$

Paso 4 Si $||\mathbf{x} - \mathbf{XO}|| < TOL$ entonces SALIDA $(x_1, \ldots, x_n)$;
(*El procedimiento fue exitoso*.)
PARE.

Paso 5 Determine $k = k + 1$.

Paso 6 Para $i = 1, \ldots, n$ determine $XO_i = x_i$.

Paso 7 SALIDA (‘número máximo de iteraciones excedido’);
(*El procedimiento no fue exitoso*.)
PARE. ■

El paso 3 del algoritmo requiere que $a_{ii} \neq 0$, para cada $i = 1, 2, \ldots, n$. Si una de las entradas a_{ii} es 0 y el sistema es no singular, se puede realizar una reorganización de las ecuaciones de tal forma que $a_{ii} = 0$. Para acelerar la convergencia, las ecuaciones se deberían reordenar de tal forma que a_{ii} resulte tan grande como sea posible. Este tema se analiza con mayor detalle más adelante en este capítulo.

Otro posible criterio para detenerse en el paso 4 es iterar hasta que

$$\frac{\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\|}{\|\mathbf{x}^{(k)}\|}$$ ■

es más pequeña que parte de la tolerancia prescrita. Para este propósito, se puede utilizar cualquier norma conveniente, la más común es la norma l_∞.

Phillip Ludwig Siedel (1821–1896) trabajó como asistente de Jacobi resolviendo problemas sobre sistemas de ecuaciones lineales que derivaron del trabajo de Gauss sobre mínimos cuadrados. En general, estas ecuaciones tienen elementos fuera de la diagonal que son mucho más pequeños que los de la diagonal, por lo que los métodos iterativos son especialmente efectivos. Las técnicas iterativas que en la actualidad se conocen como de Jacobi y Gauss-Siedel eran previamente conocidas por Gauss antes de aplicarse en esta situación; sin embargo, era frecuente que los resultados de Gauss no se difundieran ampliamente.

El método Gauss-Siedel

Una posible mejora en el algoritmo 7.1 se puede observar al reconsiderar la ecuación (7.5). Los componentes de $\mathbf{x}^{(k-1)}$ se usaban para calcular los componentes $x_i^{(k)}$ de $\mathbf{x}^{(k)}$. Pero, para $i > 1$, los componentes $x_1^{(k)}, \ldots, x_{i-1}^{(k)}$ de $\mathbf{x}^{(k)}$ ya se han calculado y se espera que sean mejores aproximaciones para las soluciones reales $x_1, \ldots, x_{i-1}$ que $x_1^{(k-1)}, \ldots, x_{i-1}^{(k-1)}$. Por lo tanto, parece razonable calcular $x_i^{(k)}$ usando estos valores calculados recientemente. Es decir, utilizar

$$x_i^{(k)} = \frac{1}{a_{ii}}\left[-\sum_{j=1}^{i-1}\left(a_{ij}x_j^{(k)}\right) - \sum_{j=i+1}^{n}\left(a_{ij}x_j^{(k-1)}\right) + b_i\right], \tag{7.8}$$

para cada $i = 1, 2, \ldots, n$, en lugar de la ecuación (7.5). Esta modificación recibe el nombre de **técnica iterativa Gauss-Siedel** y se ilustra en el siguiente ejemplo.

Ejemplo 3 Utilice la técnica iterativa Gauss-Siedel para encontrar soluciones aproximadas para

$$\begin{aligned} 10x_1 - x_2 + 2x_3 \quad &= 6, \\ -x_1 + 11x_2 - x_3 + 3x_4 &= 25, \\ 2x_1 - x_2 + 10x_3 - x_4 &= -11, \\ 3x_2 - x_3 + 8x_4 &= 15, \end{aligned}$$

que inicia con $\mathbf{x} = (0, 0, 0, 0)^t$ e iterando hasta que

$$\frac{\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\|_\infty}{\|\mathbf{x}^{(k)}\|_\infty} < 10^{-3}.$$

Solución La solución $\mathbf{x} = (1, 2, -1, 1)^t$ se aproximó usando el método de Jacobi en el ejemplo 1. Para el método Gauss-Siedel, escribimos el sistema, para cada $k = 1, 2, \ldots$ como

$$\begin{aligned} x_1^{(k)} &= \frac{1}{10}x_2^{(k-1)} - \frac{1}{5}x_3^{(k-1)} + \frac{3}{5}, \\ x_2^{(k)} &= \frac{1}{11}x_1^{(k)} + \frac{1}{11}x_3^{(k-1)} - \frac{3}{11}x_4^{(k-1)} + \frac{25}{11}, \\ x_3^{(k)} &= -\frac{1}{5}x_1^{(k)} + \frac{1}{10}x_2^{(k)} + \frac{1}{10}x_4^{(k-1)} - \frac{11}{10}, \\ x_4^{(k)} &= -\frac{3}{8}x_2^{(k)} + \frac{1}{8}x_3^{(k)} + \frac{15}{8}. \end{aligned}$$

Cuando $\mathbf{x}^{(0)} = (0, 0, 0, 0)^t$, tenemos $\mathbf{x}^{(1)} = (0.6000, 2.3272, -0.9873, 0.8789)^t$. Los valores de las iteraciones subsiguientes se muestran en la tabla 7.2.

Tabla 7.2

k	0	1	2	3	4	5
$x_1^{(k)}$	0.0000	0.6000	1.030	1.0065	1.0009	1.0001
$x_2^{(k)}$	0.0000	2.3272	2.037	2.0036	2.0003	2.0000
$x_3^{(k)}$	0.0000	−0.9873	−1.014	−1.0025	−1.0003	−1.0000
$x_4^{(k)}$	0.0000	0.8789	0.9844	0.9983	0.9999	1.0000

Puesto que

$$\frac{\|\mathbf{x}^{(5)} - \mathbf{x}^{(4)}\|_\infty}{\|\mathbf{x}^{(5)}\|_\infty} = \frac{0.0008}{2.000} = 4 \times 10^{-4},$$

$\mathbf{x}^{(5)}$ se acepta como aproximación razonable para la solución. Observe que el método de Jacobi en el ejemplo 1 requería el doble de iteraciones para la misma precisión. ■

Para escribir el método de Gauss-Siedel en forma matricial, multiplique ambos lados de la ecuación (7.8) por a_{ii} y recopile los k-ésimos términos iterados para obtener

$$a_{i1}x_1^{(k)} + a_{i2}x_2^{(k)} + \cdots + a_{ii}x_i^{(k)} = -a_{i,i+1}x_{i+1}^{(k-1)} - \cdots - a_{in}x_n^{(k-1)} + b_i,$$

para cada $i = 1, 2, \ldots, n.$ Al escribir todas las n ecuaciones nos da

$$\begin{aligned}
a_{11}x_1^{(k)} \qquad\qquad\qquad\qquad &= -a_{12}x_2^{(k-1)} - a_{13}x_3^{(k-1)} - \cdots - a_{1n}x_n^{(k-1)} + b_1,\\
a_{21}x_1^{(k)} + a_{22}x_2^{(k)} \qquad\qquad &= \qquad\qquad -a_{23}x_3^{(k-1)} - \cdots - a_{2n}x_n^{(k-1)} + b_2,\\
\vdots \qquad\qquad\qquad\qquad &\\
a_{n1}x_1^{(k)} + a_{n2}x_2^{(k)} + \cdots + a_{nn}x_n^{(k)} &= \qquad\qquad\qquad\qquad\qquad\qquad b_n;
\end{aligned}$$

con las definiciones de D, L, y U proporcionadas previamente, tenemos el método Gauss-Siedel representado por

$$(D - L)\mathbf{x}^{(k)} = U\mathbf{x}^{(k-1)} + \mathbf{b}$$

y

$$\mathbf{x}^{(k)} = (D - L)^{-1}U\mathbf{x}^{(k-1)} + (D - L)^{-1}\mathbf{b}, \quad \text{para cada} \quad k = 1, 2, \ldots. \tag{7.9}$$

Si permitimos que $T_g = (D - L)^{-1}U$ y $\mathbf{c}_g = (D - L)^{-1}\mathbf{b}$, obtenemos la técnica Gauss-Siedel de la forma

$$\mathbf{x}^{(k)} = T_g\mathbf{x}^{(k-1)} + \mathbf{c}_g. \tag{7.10}$$

Para la matriz triangular inferior $D - L$ no singular, es necesario y suficiente que $a_{ii} \neq 0$, para cada $i = 1, 2, \ldots, n$.

El algoritmo 7.2 implementa el método Gauss-Siedel.

ALGORITMO

7.2

Método iterativo Gauss-Siedel

Para resolver $A\mathbf{x} = \mathbf{b}$ dada una aproximación inicial $\mathbf{x}^{(0)}$:

ENTRADA el número de ecuaciones y valores desconocidos n; las entradas a_{ij}, $1 \leq i, j \leq n$ de la matriz A; las entradas b_i, $1 \leq i \leq n$ de $\mathbf{b}$; las entradas XO_i, $1 \leq i \leq n$ de $\mathbf{XO} = \mathbf{x}^{(0)}$; tolerancia TOL; número máximo de iteraciones N.

SALIDA la solución aproximada $x_1, \ldots, x_n$ o un mensaje que indica que se superó el número de iteraciones.

Paso 1 Determine $k = 1$.

Paso 2 Mientras ($k \leq N$) haga los pasos 3–6.

Paso 3 Para $i = 1, \ldots, n$

$$\text{Determine } x_i = \frac{1}{a_{ii}}\left[-\sum_{j=1}^{i-1} a_{ij}x_j - \sum_{j=i+1}^{n} a_{ij}XO_j + b_i\right].$$

Paso 4 Si $||\mathbf{x} - \mathbf{XO}|| < TOL$ entonces SALIDA $(x_1, \ldots, x_n)$;
(*El procedimiento fue exitoso*.)
PARE.

Paso 5 Determine $k = k + 1$.

Paso 6 Para $i = 1, \ldots, n$ determine $XO_i = x_i$.

Paso 7 SALIDA ('Número máximo de interacciones excedido');
(*El procedimiento no fue exitoso*.)
PARE.

Los comentarios que siguen al algoritmo 7.1 respecto a los criterios de reorganización e interrupción también se aplican al algoritmo 7.2 de Gauss-Siedel.

Los resultados de los ejemplos 1 y 2 parecen implicar que el método Gauss-Siedel es superior al método de Jacobi. Esto casi siempre es verdad, pero existen sistemas lineales para los que el método de Jacobi converge y el método de Gauss-Siedel no (consulte el ejercicio 9 y 10).

Métodos de iteración general

Para estudiar la convergencia de técnicas de iteración general, necesitamos analizar la fórmula

$$\mathbf{x}^{(k)} = T\mathbf{x}^{(k-1)} + \mathbf{c}, \quad \text{para cada } k = 1, 2, \ldots,$$

donde $\mathbf{x}^{(0)}$ es arbitraria. El siguiente lema y el teorema 7.17 en la página 333 dan la clave para este estudio.

Lema 7.18 Si el radio espectral satisface $\rho(T) < 1$, entonces $(I - T)^{-1}$ existe, y

$$(I - T)^{-1} = I + T + T^2 + \cdots = \sum_{j=0}^{\infty} T^j.$$

Demostración Puesto que $T\mathbf{x} = \lambda\mathbf{x}$ es verdad precisamente cuando $(I - T)\mathbf{x} = (1 - \lambda)\mathbf{x}$, tenemos λ como un eigenvalor de T precisamente cuando $1- \lambda$ es un eigenvalor de $I - T$. Pero $|\lambda| \leq \rho(T) < 1$, por lo que $\lambda = 1$ no es un eigenvalor de T y 0 no puede ser un eigenvalor de $I - T$. Por lo tanto, $(I - T)^{-1}$ existe.

Sea $S_m = I + T + T^2 + \cdots + T^m$. Entonces

$$(I - T)S_m = (1 + T + T^2 + \cdots + T^m) - (T + T^2 + \cdots + T^{m+1}) = I - T^{m+1},$$

y puesto que T es convergente, el teorema 7.17 implica que

$$\lim_{m\to\infty} (I - T)S_m = \lim_{m\to\infty} (I - T^{m+1}) = I.$$

Por lo tanto, $(I - T)^{-1} = \text{lím}_{m\to\infty} S_m = I + T + T^2 + \cdots = \sum_{j=0}^{\infty} T^j$. ■

Teorema 7.19 Para cualquier $\mathbf{x}^{(0)} \in \mathbb{R}^n$, la sucesión $\{\mathbf{x}^{(k)}\}_{k=0}^{\infty}$ definida por

$$\mathbf{x}^{(k)} = T\mathbf{x}^{(k-1)} + \mathbf{c}, \quad \text{para cada } k \geq 1, \tag{7.11}$$

converge a la solución única de $\mathbf{x} = T\mathbf{x} + \mathbf{c}$ si y sólo si $\rho(T) < 1$.

Demostración Primero suponga que $\rho(T) < 1$. Entonces,

$$\begin{aligned}\mathbf{x}^{(k)} &= T\mathbf{x}^{(k-1)} + \mathbf{c}\\ &= T(T\mathbf{x}^{(k-2)} + \mathbf{c}) + \mathbf{c}\\ &= T^2\mathbf{x}^{(k-2)} + (T + I)\mathbf{c}\\ &\ \vdots\\ &= T^k\mathbf{x}^{(0)} + (T^{k-1} + \cdots + T + I)\mathbf{c}.\end{aligned}$$

Puesto que $\rho(T) < 1$, el teorema 7.17 implica que T es convergente y

$$\lim_{k\to\infty} T^k\mathbf{x}^{(0)} = \mathbf{0}.$$

El lema 7.18 implica que

$$\lim_{k\to\infty} \mathbf{x}^{(k)} = \lim_{k\to\infty} T^k \mathbf{x}^{(0)} + \left(\sum_{j=0}^{\infty} T^j\right) \mathbf{c} = \mathbf{0} + (I-T)^{-1}\mathbf{c} = (I-T)^{-1}\mathbf{c}.$$

Por lo tanto, la sucesión $\{\mathbf{x}^{(k)}\}$ converge al vector $\mathbf{x} \equiv (I-T)^{-1}\mathbf{c}$ y $\mathbf{x} = T\mathbf{x} + \mathbf{c}$.

Para probar lo contrario, mostraremos que para cualquier $\mathbf{z} \in \mathbb{R}^n$, tenemos $\lim_{k\to\infty} T^k \mathbf{z} = \mathbf{0}$. Con el teorema 7.17, esto es equivalente a $\rho(T) < 1$.

Sea $\mathbf{z}$ un vector arbitrario y x la única solución para $\mathbf{x} = T\mathbf{x} + \mathbf{c}$. Defina $\mathbf{x}^{(0)} = \mathbf{x} - \mathbf{z}$, y, para $k \geq 1$, $\mathbf{x}^{(k)} = T\mathbf{x}^{(k-1)} + \mathbf{c}$. Entonces $\{\mathbf{x}^{(k)}\}$ converge a $\mathbf{x}$. También,

$$\mathbf{x} - \mathbf{x}^{(k)} = (T\mathbf{x} + \mathbf{c}) - \left(T\mathbf{x}^{(k-1)} + \mathbf{c}\right) = T\left(\mathbf{x} - \mathbf{x}^{(k-1)}\right),$$

por lo que

$$\mathbf{x} - \mathbf{x}^{(k)} = T\left(\mathbf{x} - \mathbf{x}^{(k-1)}\right) = T^2\left(\mathbf{x} - \mathbf{x}^{(k-2)}\right) = \cdots = T^k\left(\mathbf{x} - \mathbf{x}^{(0)}\right) = T^k\mathbf{z}.$$

Por lo tanto, $\lim_{k\to\infty} T^k\mathbf{z} = \lim_{k\to\infty} T^k\left(\mathbf{x} - \mathbf{x}^{(0)}\right) = \lim_{k\to\infty}\left(\mathbf{x} - \mathbf{x}^{(k)}\right) = \mathbf{0}$.

Pero $\mathbf{z} \in \mathbb{R}^n$ era arbitrario, por lo que mediante el teorema 7.17, T es convergente y $\rho(T) < 1$. ■

La prueba del siguiente corolario es similar a las pruebas en el corolario 2.5 en la página 47. Se considera en el ejercicio 18.

Corolario 7.20 Si $\|T\| < 1$ para cualquier norma matricial normal y $\mathbf{c}$ es un vector determinado, entonces la sucesión $\{\mathbf{x}^{(k)}\}_{k=0}^{\infty}$ definida por $\mathbf{x}^{(k)} = T\mathbf{x}^{(k-1)} + \mathbf{c}$ converge, para cualquier $\mathbf{x}^{(0)} \in \mathbb{R}^n$, para un vector $\mathbf{x} \in \mathbb{R}^n$, con $\mathbf{x} = T\mathbf{x} + \mathbf{c}$, y las siguientes cotas de error se mantienen:

i) $\|\mathbf{x} - \mathbf{x}^{(k)}\| \leq \|T\|^k \|\mathbf{x}^{(0)} - \mathbf{x}\|$; **ii)** $\|\mathbf{x} - \mathbf{x}^{(k)}\| \leq \frac{\|T\|^k}{1-\|T\|}\|\mathbf{x}^{(1)} - \mathbf{x}^{(0)}\|$. ■

Hemos observado que las técnicas de Jacobi y Gauss-Siedel se pueden escribir como

$$\mathbf{x}^{(k)} = T_j\mathbf{x}^{(k-1)} + \mathbf{c}_j \quad \text{y} \quad \mathbf{x}^{(k)} = T_g\mathbf{x}^{(k-1)} + \mathbf{c}_g$$

usando las matrices

$$T_j = D^{-1}(L+U) \quad \text{y} \quad T_g = (D-L)^{-1}U.$$

Si $\rho(T_j)$ o $\rho(T_g)$ es menor a 1, entonces la sucesión correspondiente $\{\mathbf{x}^{(k)}\}_{k=0}^{\infty}$ convergerá a la solución $\mathbf{x}$ de $A\mathbf{x} = \mathbf{b}$. Por ejemplo, el esquema de Jacobi tiene

$$\mathbf{x}^{(k)} = D^{-1}(L+U)\mathbf{x}^{(k-1)} + D^{-1}\mathbf{b},$$

y si $\{\mathbf{x}^{(k)}\}_{k=0}^{\infty}$ converge a $\mathbf{x}$, entonces

$$\mathbf{x} = D^{-1}(L+U)\mathbf{x} + D^{-1}\mathbf{b}.$$

Esto implica que

$$D\mathbf{x} = (L+U)\mathbf{x} + \mathbf{b} \quad \text{y} \quad (D-L-U)\mathbf{x} = \mathbf{b}.$$

Puesto que $D - L - U = A$, la solución $\mathbf{x}$ satisface $A\mathbf{x} = \mathbf{b}$.

Ahora podemos proporcionar condiciones de suficiencia verificadas fácilmente para convergencia de los métodos de Jacobi y de Gauss-Siedel. (Para probar la convergencia para el esquema de Jacobi, consulte el ejercicio 17, y para el esquema de Gauss-Siedel, consulte [Or2], p. 120).

Teorema 7.21 Si A es estrictamente diagonalmente dominante, entonces para cualquier selección de $\mathbf{x}^{(0)}$, tanto los métodos de Gauss-Siedel y de Jacobi dan sucesiones $\{\mathbf{x}^{(k)}\}_{k=0}^{\infty}$ que convergen a la solución única de $A\mathbf{x} = \mathbf{b}$. ■

La relación de la rapidez de convergencia para el radio espectral de la matriz de iteración T se puede observar a partir del corolario 7.20. Las desigualdades se mantienen para cualquier norma matricial natural, por lo que sigue la declaración después del teorema 7.15 en la página 332 que

$$\|\mathbf{x}^{(k)} - \mathbf{x}\| \approx \rho(T)^k \|\mathbf{x}^{(0)} - \mathbf{x}\|. \tag{7.12}$$

Por lo tanto, nos gustaría seleccionar la técnica iterativa con $\rho(T) < 1$ mínima para un sistema particular $A\mathbf{x} = \mathbf{b}$. No existen resultados generales para decir cuál de las dos técnicas, Jacobi o Gauss-Siedel, será más exitosa para cualquier sistema lineal arbitraria. En casos especiales, sin embargo, la respuesta es conocida, como se demuestra en el siguiente teorema. La prueba de este resultado se puede encontrar en [y], p. 120–127.

Teorema 7.22 **(Stein-Rosenberg)**

Si $a_{ij} \leq 0$, para cada $i \neq j$, y $a_{ii} > 0$, para cada $i = 1, 2, \ldots, n$, entonces una y sólo una de las siguientes declaraciones es válida:

i) $0 \leq \rho(T_g) < \rho(T_j) < 1$; **ii)** $1 < \rho(T_j) < \rho(T_g)$;
iii) $\rho(T_j) = \rho(T_g) = 0$; **iv)** $\rho(T_j) = \rho(T_g) = 1$. ■

Para el caso especial descrito en el teorema 7.22, observamos, a partir de la parte i), que cuando un método proporciona convergencia, entonces provee convergencia y el método de Gauss-Siedel converge más rápido que el método de Jacobi. La parte ii) indica que cuando un método diverge, entonces ambos divergen y la divergencia es más pronunciada para el método Gauss-Siedel.

La sección Conjunto de ejercicios 7.3 está disponible en línea. Encuentre la ruta de acceso en las páginas preliminares.

7.4 Técnicas de relajación para resolver sistemas lineales

En la sección 7.3 observamos que la tasa de convergencia de una técnica iterativa depende del radio espectral de la matriz relacionada con el método. Una forma de seleccionar un procedimiento para convergencia acelerada es seleccionar un método cuya matriz relacionada tiene un radio espectral mínimo. Antes de describir un procedimiento para seleccionar dicho método, necesitamos introducir medios nuevos para medir la cantidad por la que una aproximación a la solución de un sistema lineal difiere de la verdadera solución del sistema. El método utiliza el vector descrito en la siguiente definición.

Definición 7.23 Suponga que $\tilde{\mathbf{x}} \in \mathbb{R}^n$ es una aproximación a la solución del sistema lineal definido por $A\mathbf{x} = \mathbf{b}$. El vector residual para $\tilde{\mathbf{x}}$ respecto a este sistema es $\mathbf{r} = \mathbf{b} - A\tilde{\mathbf{x}}$. ■

La palabra "residual" significa lo que sobra, por lo que es un nombre adecuado para este vector.

En procedimientos como los métodos de Jacobi o Gauss-Siedel un vector residual está relacionado con cada cálculo de un componente aproximado para el vector solución. El verdadero objetivo es generar una sucesión de aproximaciones que causarán que los vectores residuales converjan rápidamente a cero. Suponga que tenemos

$$\mathbf{r}_i^{(k)} = \left(r_{1i}^{(k)}, r_{2i}^{(k)}, \ldots, r_{ni}^{(k)}\right)^t$$

denota el vector residual para el método Gauss-Siedel correspondiente al vector solución aproximado $\mathbf{x}_i^{(k)}$ definido por

$$\mathbf{x}_i^{(k)} = \left(x_1^{(k)}, x_2^{(k)}, \ldots, x_{i-1}^{(k)}, x_i^{(k-1)}, \ldots, x_n^{(k-1)}\right)^t.$$

El m-ésimo componente de $\mathbf{r}_i^{(k)}$ es

$$r_{mi}^{(k)} = b_m - \sum_{j=1}^{i-1} a_{mj} x_j^{(k)} - \sum_{j=i}^{n} a_{mj} x_j^{(k-1)}, \tag{7.13}$$

o, de manera equivalente,

$$r_{mi}^{(k)} = b_m - \sum_{j=1}^{i-1} a_{mj} x_j^{(k)} - \sum_{j=i+1}^{n} a_{mj} x_j^{(k-1)} - a_{mi} x_i^{(k-1)},$$

para cada $m = 1, 2, \ldots, n$.

En particular, el i-ésimo componente de $\mathbf{r}_i^{(k)}$ es

$$r_{ii}^{(k)} = b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(k)} - \sum_{j=i+1}^{n} a_{ij} x_j^{(k-1)} - a_{ii} x_i^{(k-1)},$$

por lo que

$$a_{ii} x_i^{(k-1)} + r_{ii}^{(k)} = b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(k)} - \sum_{j=i+1}^{n} a_{ij} x_j^{(k-1)}. \tag{7.14}$$

Sin embargo, recuerde que el método Gauss-Siedel, $x_i^{(k)}$ se selecciona como

$$x_i^{(k)} = \frac{1}{a_{ii}} \left[b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(k)} - \sum_{j=i+1}^{n} a_{ij} x_j^{(k-1)} \right], \tag{7.15}$$

por lo que la ecuación (7.14) se puede reescribir como

$$a_{ii} x_i^{(k-1)} + r_{ii}^{(k)} = a_{ii} x_i^{(k)}.$$

Por consiguiente, el método Gauss-Siedel se puede caracterizar seleccionando $x_i^{(k)}$ para satisfacer

$$x_i^{(k)} = x_i^{(k-1)} + \frac{r_{ii}^{(k)}}{a_{ii}}. \tag{7.16}$$

Podemos derivar otra conexión entre los vectores residuales y la técnica de Gauss-Siedel. Considere el vector residual $\mathbf{r}_{i+1}^{(k)}$, asociado con el vector $\mathbf{x}_{i+1}^{(k)} = (x_1^{(k)}, \ldots, x_i^{(k)}, x_{i+1}^{(k-1)}, \ldots, x_n^{(k-1)})^t$. Mediante la ecuación (7.13), el i-ésimo componente de $\mathbf{r}_{i+1}^{(k)}$ es

$$\begin{aligned} r_{i,i+1}^{(k)} &= b_i - \sum_{j=1}^{i} a_{ij} x_j^{(k)} - \sum_{j=i+1}^{n} a_{ij} x_j^{(k-1)} \\ &= b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(k)} - \sum_{j=i+1}^{n} a_{ij} x_j^{(k-1)} - a_{ii} x_i^{(k)}. \end{aligned}$$

Solución para cada $k = 1, 2, \ldots$, las ecuaciones para el método de Gauss-Siedel son

$$x_1^{(k)} = -0.75x_2^{(k-1)} + 6,$$
$$x_2^{(k)} = -0.75x_1^{(k)} + 0.25x_3^{(k-1)} + 7.5,$$
$$x_3^{(k)} = 0.25x_2^{(k)} - 6,$$

y las ecuaciones para el método SOR con $\omega = 1.25$ son

$$x_1^{(k)} = -0.25x_1^{(k-1)} - 0.9375x_2^{(k-1)} + 7.5,$$
$$x_2^{(k)} = -0.9375x_1^{(k)} - 0.25x_2^{(k-1)} + 0.3125x_3^{(k-1)} + 9.375,$$
$$x_3^{(k)} = 0.3125x_2^{(k)} - 0.25x_3^{(k-1)} - 7.5.$$

Las primeras siete iteraciones para cada método se listan en las tablas 7.3 y 7.4. Para que las iteraciones sean precisas para siete lugares decimales, el método de Gauss-Siedel requiere 34 iteraciones, en comparación con las 14 del método SOR con $\omega = 1.25$. ■

Tabla 7.3

k	0	1	2	3	4	5	6	7
$x_1^{(k)}$	1	5.250000	3.1406250	3.0878906	3.0549316	3.0343323	3.0214577	3.0134110
$x_2^{(k)}$	1	3.812500	3.8828125	3.9267578	3.9542236	3.9713898	3.9821186	3.9888241
$x_3^{(k)}$	1	−5.046875	−5.0292969	−5.0183105	−5.0114441	−5.0071526	−5.0044703	−5.0027940

Tabla 7.4

k	0	1	2	3	4	5	6	7
$x_1^{(k)}$	1	6.3125000	2.6223145	3.1333027	2.9570512	3.0037211	2.9963276	3.0000498
$x_2^{(k)}$	1	3.5195313	3.9585266	4.0102646	4.0074838	4.0029250	4.0009262	4.0002586
$x_3^{(k)}$	1	−6.6501465	−4.6004238	−5.0966863	−4.9734897	−5.0057135	−4.9982822	−5.0003486

Una pregunta obvia es cómo se selecciona el valor adecuado de ω cuando se usa el método SOR. A pesar de que no se conoce una respuesta completa a esta pregunta para el sistema lineal $n \times n$, los siguientes resultados se pueden utilizar en ciertas situaciones importantes.

Teorema 7.24 **(Kahan)**

Si $a_{ii} \neq 0$, para cada $i = 1, 2, \ldots, n$, entonces $\rho(T_\omega) \geq |\omega - 1|$. Esto implica que el método SOR puede converger sólo si $0 < \omega < 2$. ■

La demostración de este teorema se considera en el ejercicio 13. La de los siguientes dos resultados se puede encontrar en [Or2], p. 123-133. Estos resultados se usarán en el capítulo 12.

Teorema 7.25 **(Ostrowski-Reich)**

Si A es una matriz definida positiva y $0 < \omega < 2$ entonces el método SOR converge para cualquier opción de vector aproximado inicial $\mathbf{x}^{(0)}$. ■

Teorema 7.26 Si A es definida positiva y tridiagonal, entonces $\rho(T_g) = [\rho(T_j)]^2 < 1$, y la selección óptima de ω, para el método SOR es

$$\omega = \frac{2}{1 + \sqrt{1 - [\rho(T_j)]^2}}.$$

Con esta selección de ω, tenemos $\rho(T_\omega) = \omega - 1$. ■

Ejemplo 2 Encuentre la selección óptima de ω, para el método SOR para la matriz

$$A = \begin{bmatrix} 4 & 3 & 0 \\ 3 & 4 & -1 \\ 0 & -1 & 4 \end{bmatrix}.$$

Solución Esta matriz es claramente triangular, por lo que podemos aplicar el resultado del teorema 7.26 si también podemos mostrar que es definida positiva. Puesto que la matriz es simétrica, el teorema 6.25 en la página 310 establece que es definida positiva si y sólo si todas sus primeras submatrices principales tienen determinantes positivos. Esto se observa fácilmente es este caso porque

$$\det(A) = 24, \ \det\left(\begin{bmatrix} 4 & 3 \\ 3 & 4 \end{bmatrix}\right) = 7, \ \text{y} \det\ ([4]) = 4.$$

Puesto que

$$T_j = D^{-1}(L+U) = \begin{bmatrix} \frac{1}{4} & 0 & 0 \\ 0 & \frac{1}{4} & 0 \\ 0 & 0 & \frac{1}{4} \end{bmatrix} \begin{bmatrix} 0 & -3 & 0 \\ -3 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} = \begin{bmatrix} 0 & -0.75 & 0 \\ -0.75 & 0 & 0.25 \\ 0 & 0.25 & 0 \end{bmatrix},$$

tenemos

$$T_j - \lambda I = \begin{bmatrix} -\lambda & -0.75 & 0 \\ -0.75 & -\lambda & 0.25 \\ 0 & 0.25 & -\lambda \end{bmatrix},$$

por lo que

$$\det(T_j - \lambda I) = -\lambda(\lambda^2 - 0.625).$$

Por lo tanto,

$$\rho(T_j) = \sqrt{0.625}$$

y

$$\omega = \frac{2}{1 + \sqrt{1 - [\rho(T_j)]^2}} = \frac{2}{1 + \sqrt{1 - 0.625}} \approx 1.24.$$

Esto explica la rápida convergencia obtenida en el ejemplo 1 cuando se utiliza $\omega = 1.25$. ■

Cerramos esta sección con el algoritmo 7.3 para el método SOR.

ALGORITMO
7.3

SOR

Para resolver $A\mathbf{x} = \mathbf{b}$ dado el parámetro ω y una aproximación inicial $\mathbf{x}^{(0)}$:

ENTRADA el número de ecuaciones y valores desconocidos n; las entradas a_{ij}, $1 \leq i, j \leq n$ de la matriz A; las entradas b_i, $1 \leq i \leq n$, de $\mathbf{b}$; las entradas XO_i, $1 \leq i \leq n$, de $\mathbf{XO} = \mathbf{x}^{(0)}$; el parámetro ω; tolerancia TOL; el número máximo de iteraciones N.

SALIDA la solución aproximada $x_1, \ldots, x_n$ o un mensaje que indique que se superó el número de iteraciones.

Paso 1 Determine $k = 1$.

Paso 2 Mientras $(k \leq N)$ haga los pasos 3–6.

Paso 3 Para $i = 1, \ldots, n$

determine $x_i = (1-\omega)XO_i +$

$$\frac{1}{a_{ii}}\left[\omega\left(-\sum_{j=1}^{i-1} a_{ij}x_j - \sum_{j=i+1}^{n} a_{ij}XO_j + b_i\right)\right].$$

Paso 4 Si $||\mathbf{x} - \mathbf{XO}|| < TOL$ entonces SALIDA $(x_1, \ldots, x_n)$;
(*El procedimiento fue exitoso*.)
PARE.

Paso 5 Determine $k = k + 1$.

Paso 6 Para $i = 1, \ldots, n$ determine $XO_i = x_i$.

Paso 7 SALIDA ('Número máximo de interacciones alcanzado');
(*El procedimiento no fue exitoso*.)
PARE. ■

La sección Conjunto de ejercicios 7.4 está disponible en línea. Encuentre la ruta de acceso en las páginas preliminares.

7.5 Cotas de error y refinamiento iterativo

Parece intuitivamente razonable que si $\tilde{\mathbf{x}}$ es una aproximación a la solución $\mathbf{x}$ de $A\mathbf{x} = \mathbf{b}$ y el vector residual $\mathbf{r} = \mathbf{b} - A\tilde{\mathbf{x}}$ tiene la propiedad de que $\|\mathbf{r}\|$ es pequeña, entonces $\|\mathbf{x} - \tilde{\mathbf{x}}\|$ también sería pequeña. A menudo, éste es el caso, pero ciertos sistemas, a menudo presentes en la práctica carecen de esta propiedad.

Ejemplo 1 El sistema lineal $A\mathbf{x} = \mathbf{b}$ dado por

$$\begin{bmatrix} 1 & 2 \\ 1.0001 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 3 \\ 3.0001 \end{bmatrix}$$

tiene la solución única $\mathbf{x} = (1, 1)^t$. Determine el vector residual para la aproximación deficiente $\tilde{\mathbf{x}} = (3, -0.0001)^t$.

Solución Tenemos

$$\mathbf{r} = \mathbf{b} - A\tilde{\mathbf{x}} = \begin{bmatrix} 3 \\ 3.0001 \end{bmatrix} - \begin{bmatrix} 1 & 2 \\ 1.0001 & 2 \end{bmatrix} \begin{bmatrix} 3 \\ -0.0001 \end{bmatrix} = \begin{bmatrix} 0.0002 \\ 0 \end{bmatrix},$$

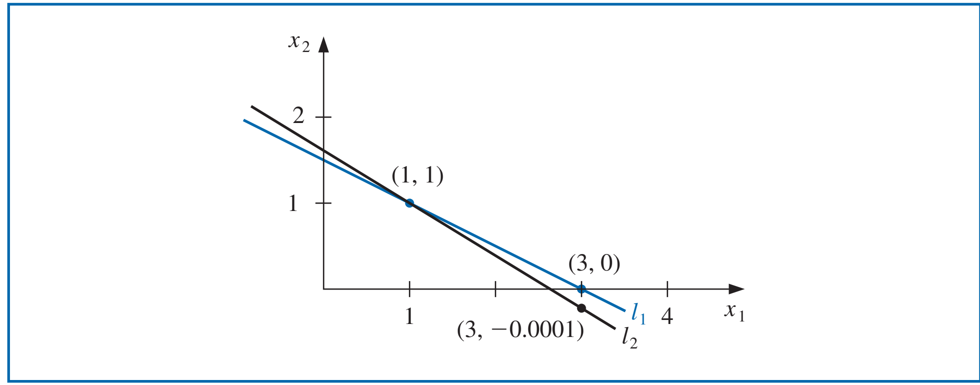
por lo que $\|\mathbf{r}\|_\infty = 0.0002$. A pesar de que la norma del vector residual es pequeña, la aproximación $\tilde{\mathbf{x}} = (3, -0.0001)^t$ es obviamente bastante deficiente; de hecho, $\|\mathbf{x} - \tilde{\mathbf{x}}\|_\infty = 2$. ■

La dificultad en el ejemplo 1 se explica con facilidad al observar que la solución del sistema representa la intersección de las rectas

$$l_1: \quad x_1 + 2x_2 = 3 \quad \text{y} \quad l_2: \quad 1.0001x_1 + 2x_2 = 3.0001.$$

El punto $(3, -0.0001)$ se encuentra en l_2, y las rectas son casi paralelas. Esto implica que $(3, -0.0001)$ también se encuentra cerca de l_1, aunque difiere significativamente de la solución del sistema, determinada por el punto de intersección $(1, 1)$. (Consulte la figura 7.7.)

Figura 7.7



El ejemplo 1 se construyó claramente para mostrar las dificultades que pueden surgir, y de hecho, surgen. Si las rectas no coinciden por completo, esperaríamos que un vector residual pequeño implique una aproximación precisa.

En la situación general no podemos depender de la geometría del sistema para proporcionar una indicación de cuándo pueden surgir los problemas. Sin embargo, podemos obtener esta información al considerar las normas de la matriz A y su inversa.

Teorema 7.27 Suponga que $\tilde{\mathbf{x}}$ es una aproximación a la solución de $A\mathbf{x} = \mathbf{b}$, A es una matriz no singular y $\mathbf{r}$ es el vector residual para $\tilde{\mathbf{x}}$. Entonces, para cualquier norma natural,

$$\|\mathbf{x} - \tilde{\mathbf{x}}\| \leq \|\mathbf{r}\| \cdot \|A^{-1}\|,$$

y si $\mathbf{x} \neq \mathbf{0}$ y $\mathbf{b} \neq \mathbf{0}$,

$$\frac{\|\mathbf{x} - \tilde{\mathbf{x}}\|}{\|\mathbf{x}\|} \leq \|A\| \cdot \|A^{-1}\| \frac{\|\mathbf{r}\|}{\|\mathbf{b}\|}. \tag{7.20}$$

Demostración Puesto que $\mathbf{r} = \mathbf{b} - A\tilde{\mathbf{x}} = A\mathbf{x} - A\tilde{\mathbf{x}}$ y A es no singular, tenemos $\mathbf{x} - \tilde{\mathbf{x}} = A^{-1}\mathbf{r}$. El corolario 7.10 en la página 326 implica que

$$\|\mathbf{x} - \tilde{\mathbf{x}}\| = \|A^{-1}\mathbf{r}\| \leq \|A^{-1}\| \cdot \|\mathbf{r}\|.$$

Además, puesto que $\mathbf{b} = A\mathbf{x}$, tenemos $\|\mathbf{b}\| \leq \|A\| \cdot \|\mathbf{x}\|$. Por lo que $1/\|\mathbf{x}\| \leq \|A\|/\|\mathbf{b}\|$ y

$$\frac{\|\mathbf{x} - \tilde{\mathbf{x}}\|}{\|\mathbf{x}\|} \leq \frac{\|A\| \cdot \|A^{-1}\|}{\|\mathbf{b}\|} \|\mathbf{r}\|.$$ ■

Números de condición

Las desigualdades en el teorema 7.27 implican que $\|A^{-1}\|$ y $\|A\| \cdot \|A^{-1}\|$ proveen una indicación de la conexión entre el vector residual y la precisión de la aproximación. En general el error relativo $\|\mathbf{x} - \tilde{\mathbf{x}}\|/\|\mathbf{x}\|$ es de mayor interés, y, mediante la desigualdad (7.20), este error está acotado por el producto de $\|A\| \cdot \|A^{-1}\|$ con el residuo relativo de esta aproximación, $\|\mathbf{r}\|/\|\mathbf{b}\|$. Cualquier norma conveniente se puede usar para esta aproximación; el único requisito es que se use constantemente de principio a fin.

Definición 7.28 El **número de condición** de la matriz no singular A relativo a la norma $\|\cdot\|$ es

$$K(A) = \|A\| \cdot \|A^{-1}\|.$$ ■

Con esta notación, las desigualdades en el teorema 7.27 se convierten en

$$\|\mathbf{x} - \tilde{\mathbf{x}}\| \leq K(A) \frac{\|\mathbf{r}\|}{\|A\|}$$

y

$$\frac{\|\mathbf{x}-\tilde{\mathbf{x}}\|}{\|\mathbf{x}\|} \leq K(A)\frac{\|\mathbf{r}\|}{\|\mathbf{b}\|}.$$

Para cualquier matriz A no singular y norma natural $\|\cdot\|$,

$$1 = \|I\| = \|A \cdot A^{-1}\| \leq \|A\| \cdot \|A^{-1}\| = K(A).$$

Una matriz A está **bien condicionada** si $K(A)$ está cerca de 1 y está **mal condicionada** cuando $K(A)$ es significativamente mayor que 1. El condicionamiento en este contexto se refiere a la seguridad relativa de que un vector residual pequeño implica una solución aproximada correspondientemente precisa.

Ejemplo 2 Determine el número de condición para la matriz

$$A = \begin{bmatrix} 1 & 2 \\ 1.0001 & 2 \end{bmatrix}.$$

Solución En el ejemplo 1 observamos que la misma aproximación deficiente $(3, -0.0001)^t$ para la solución exacta $(1, 1)^t$ tenía un vector residual con norma pequeña, por lo que deberíamos esperar que el número de condición de A sea grande. Tenemos $\|A\|_\infty = \text{máx}\{|1| + |2|, |1.001| + |2|\} = 3.0001$, que no se podría considerar grande. Sin embargo,

$$A^{-1} = \begin{bmatrix} -10000 & 10000 \\ 5000.5 & -5000 \end{bmatrix}, \quad \text{así} \quad \|A^{-1}\|_\infty = 20000,$$

y para la norma de infinidad, $K(A) = (20000)(3.0001) = 60002$. El tamaño del número de condición para este ejemplo debería evitar que tomáramos decisiones apresuradas con base en el residuo de una aproximación. ■

A pesar de que un número de condición de una matriz depende solamente de las normas de la matriz y su inversa, el cálculo de la inversa está sujeto a error de redondeo y depende de la precisión con la que se realizan los cálculos. Si las operaciones implican aritmética con t dígitos de precisión, el número de condición aproximada para la matriz A es la norma de la matriz multiplicada por la norma de la aproximación para la inversa de A, que se obtiene a través de aritmética de t dígitos. De hecho, este número de condición también depende del método utilizado para calcular la inversa de A. Además, debido al número de cálculos necesarios para calcular la inversa, necesitamos ser capaces de calcular el número de condición sin determinar directamente la inversa.

Si suponemos que la solución aproximada para el sistema lineal $A\mathbf{x} = \mathbf{b}$ se determina por medio de la aritmética de t dígitos y la eliminación gaussiana, es posible mostrar (consulte [FM], p. 45-47) que el vector $\mathbf{r}$ para la aproximación $\tilde{\mathbf{x}}$ tiene

$$\|\mathbf{r}\| \approx 10^{-t}\|A\| \cdot \|\tilde{\mathbf{x}}\|. \tag{7.21}$$

A partir de esta aproximación se puede obtener un cálculo para el número de condición efectivo en aritmética de t dígitos sin necesidad de invertir la matriz A. En la actualidad esta aproximación supone que todas las operaciones aritméticas en la técnica de eliminación gaussiana se realizan usando la aritmética de t dígitos, pero que las operaciones necesarias para determinar el residuo se realizan en aritmética de doble precisión (es decir, $2t$ dígitos). Esta técnica no se suma significativamente al esfuerzo computacional y elimina la mayor parte de la pérdida de precisión implicada con la resta de los números casi iguales que se presentan en el cálculo del residuo.

La aproximación para el número de condición $K(A)$ con t dígitos proviene de la consideración del sistema lineal

$$A\mathbf{y} = \mathbf{r}.$$

La solución para este sistema se puede aproximar fácilmente porque los multiplicadores para el método de eliminación gaussiana ya se han calculado. Por ello, A se puede factorizar de la forma $P^t LU$, como se describe en la sección 5 del capítulo 6. De hecho, $\tilde{\mathbf{y}}$, la solución aproximada de $A\mathbf{y} = \mathbf{r}$, satisface

$$\tilde{\mathbf{y}} \approx A^{-1}\mathbf{r} = A^{-1}(\mathbf{b} - A\tilde{\mathbf{x}}) = A^{-1}\mathbf{b} - A^{-1}A\tilde{\mathbf{x}} = \mathbf{x} - \tilde{\mathbf{x}}, \tag{7.22}$$

y

$$\mathbf{x} \approx \tilde{\mathbf{x}} + \tilde{\mathbf{y}}.$$

Por lo que $\tilde{\mathbf{y}}$ es un cálculo del error producido cuando $\tilde{\mathbf{x}}$ se aproxima a la solución $\mathbf{x}$ del sistema original. Las ecuaciones (7.21) y (7.22) implican que

$$\|\tilde{\mathbf{y}}\| \approx \|\mathbf{x} - \tilde{\mathbf{x}}\| = \|A^{-1}\mathbf{r}\| \leq \|A^{-1}\| \cdot \|\mathbf{r}\| \approx \|A^{-1}\|(10^{-t}\|A\| \cdot \|\tilde{\mathbf{x}}\|) = 10^{-t}\|\tilde{\mathbf{x}}\|K(A).$$

Esto nos da una aproximación para el número de condición que participa en la solución del sistema $A\mathbf{x} = \mathbf{b}$ usando eliminación gaussiana y el tipo de t dígitos de la aritmética que acabamos de describir:

$$K(A) \approx \frac{\|\tilde{\mathbf{y}}\|}{\|\tilde{\mathbf{x}}\|}10^t. \tag{7.23}$$

Ilustración El sistema lineal dado por

$$\begin{bmatrix} 3.3330 & 15920 & -10.333 \\ 2.2220 & 16.710 & 9.6120 \\ 1.5611 & 5.1791 & 1.6852 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 15913 \\ 28.544 \\ 8.4254 \end{bmatrix}$$

tiene la solución exacta $\mathbf{x} = (1, 1, 1)^t$.

Mediante eliminación gaussiana y aritmética de redondeo de cinco dígitos conduce sucesivamente a las matrices aumentadas

$$\begin{bmatrix} 3.3330 & 15920 & -10.333 & 15913 \\ 0 & -10596 & 16.501 & 10580 \\ 0 & -7451.4 & 6.5250 & -7444.9 \end{bmatrix}$$

y

$$\begin{bmatrix} 3.3330 & 15920 & -10.333 & 15913 \\ 0 & -10596 & 16.501 & -10580 \\ 0 & 0 & -5.0790 & -4.7000 \end{bmatrix}.$$

La solución aproximada para este sistema es

$$\tilde{\mathbf{x}} = (1.2001, 0.99991, 0.92538)^t.$$

El vector residual correspondiente a $\tilde{\mathbf{x}}$ se calcula con precisión doble como

$$\begin{aligned} \mathbf{r} &= \mathbf{b} - A\tilde{\mathbf{x}} \\ &= \begin{bmatrix} 15913 \\ 28.544 \\ 8.4254 \end{bmatrix} - \begin{bmatrix} 3.3330 & 15920 & -10.333 \\ 2.2220 & 16.710 & 9.6120 \\ 1.5611 & 5.1791 & 1.6852 \end{bmatrix} \begin{bmatrix} 1.2001 \\ 0.99991 \\ 0.92538 \end{bmatrix} \\ &= \begin{bmatrix} 15913 \\ 28.544 \\ 8.4254 \end{bmatrix} - \begin{bmatrix} 15913.00518 \\ 28.26987086 \\ 8.611560367 \end{bmatrix} = \begin{bmatrix} -0.00518 \\ 0.27412914 \\ -0.186160367 \end{bmatrix}, \end{aligned}$$

Por lo que

$$\|\mathbf{r}\|_\infty = 0.27413.$$

El cálculo para el número de condición provisto en el análisis anterior se obtiene al resolver primero el sistema $A\mathbf{y} = \mathbf{r}$ para $\tilde{\mathbf{y}}$:

$$\begin{bmatrix} 3.3330 & 15920 & -10.333 \\ 2.2220 & 16.710 & 9.6120 \\ 1.5611 & 5.1791 & 1.6852 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} -0.00518 \\ 0.27413 \\ -0.18616 \end{bmatrix}.$$

Esto implica que $\tilde{\mathbf{y}} = (-0.20008, 8.9987 \times 10^{-5}, 0.074607)^t$. Mediante el cálculo en la ecuación (7.23) da

$$K(A) \approx \frac{\|\tilde{\mathbf{y}}\|_\infty}{\|\tilde{\mathbf{x}}\|_\infty} 10^5 = \frac{0.20008}{1.2001} 10^5 = 16672. \tag{7.24}$$

Para determinar el número de condición *exacto* de A, primero debemos encontrar A^{-1}. Usando aritmética de cinco dígitos para los cálculos obtenemos la aproximación

$$A^{-1} \approx \begin{bmatrix} -1.1701 \times 10^{-4} & -1.4983 \times 10^{-1} & 8.5416 \times 10^{-1} \\ 6.2782 \times 10^{-5} & 1.2124 \times 10^{-4} & -3.0662 \times 10^{-4} \\ -8.6631 \times 10^{-5} & 1.3846 \times 10^{-1} & -1.9689 \times 10^{-1} \end{bmatrix}.$$

El teorema 7.11 en la página 328 implica que $\|A^{-1}\|_\infty = 1.0041$ y $\|A\|_\infty = 15934$.

Como consecuencia, la matriz mal condicionada A tiene

$$K(A) = (1.0041)(15934) = 15999.$$

El cálculo en la ecuación (7.24) es bastante cercano a $K(A)$ y requiere considerablemente menos esfuerzo computacional.

Puesto que se conoce la solución real $\mathbf{x} = (1, 1, 1)^t$ para este sistema, podemos calcular tanto

$$\|\mathbf{x} - \tilde{\mathbf{x}}\|_\infty = 0.2001 \quad \text{y} \quad \frac{\|\mathbf{x} - \tilde{\mathbf{x}}\|_\infty}{\|\mathbf{x}\|_\infty} = \frac{0.2001}{1} = 0.2001.$$

Las cotas de error dadas en el teorema 7.27 para estos valores son

$$\|\mathbf{x} - \tilde{\mathbf{x}}\|_\infty \le K(A) \frac{\|\mathbf{r}\|_\infty}{\|A\|_\infty} = \frac{(15999)(0.27413)}{15934} = 0.27525$$

y

$$\frac{\|\mathbf{x} - \tilde{\mathbf{x}}\|_\infty}{\|\mathbf{x}\|_\infty} \le K(A) \frac{\|\mathbf{r}\|_\infty}{\|\mathbf{b}\|_\infty} = \frac{(15999)(0.27413)}{15913} = 0.27561.$$

■

Refinamiento iterativo

En la ecuación (7.22) utilizamos el cálculo $\tilde{\mathbf{y}} \approx \mathbf{x} - \tilde{\mathbf{x}}$, donde $\tilde{\mathbf{y}}$ es la solución aproximada para el sistema $A\mathbf{y} = \mathbf{r}$. En general, $\tilde{\mathbf{x}}+\tilde{\mathbf{y}}$ es una aproximación más precisa del sistema lineal $A\mathbf{x} = \mathbf{b}$ que la original $\tilde{\mathbf{x}}$. El método que usa esta suposición recibe el nombre de **refinamiento iterativo**, o *mejora iterativa*, y consiste en realizar iteraciones sobre el sistema cuyo lado derecho es el vector residual para aproximaciones sucesivas hasta obtener resultados precisos satisfactorios.

Si se aplica el proceso mediante aritmética de t dígitos y si $K_\infty(A) \approx 10^q$, entonces después de k iteraciones de refinamiento iterativo, la solución tiene aproximadamente el dígito más pequeño de t y $k(t - q)$ dígitos correctos. Si el sistema está bien condicionado, una o dos iteraciones indicarán que la solución es precisa. Existe la posibilidad de mejora significativa en sistemas mal condicionados a menos que la matriz A también esté tan mal condicionada que $K_\infty(A) > 10^t$. En esa situación, se usaría la precisión incrementada para los cálculos. El algoritmo 7.4 implementa el método de refinamiento iterativo.

ALGORITMO **7.4**

Refinamiento iterativo

Para aproximar la solución del sistema lineal $A\mathbf{x} = \mathbf{b}$:

ENTRADA el número de ecuaciones y valores desconocidos n; las entradas a_{ij}, $1 \leq i, j \leq n$ de la matriz A; las entradas b_i, $1 \leq i \leq n$ de $\mathbf{b}$; el número máximo de iteraciones N; tolerancia TOL; número de dígitos de precisión t.

SALIDA la aproximación $\mathbf{xx} = (xx_i, \ldots, xx_n)^t$ o un mensaje de que el número de iteraciones fue excedido y una aproximación $COND$ para $K_\infty(A)$.

Paso 0 Resuelva el sistema $A\mathbf{x} = \mathbf{b}$ para $x_1, \ldots, x_n$ usando eliminación gaussiana al guardar los multiplicadores m_{ji}, $j = i + 1, i + 2, \ldots, n$, $i = 1, 2, \ldots, n - 1$ y observar los intercambios de fila.

Paso 1 Determine $k = 1$.

Paso 2 Mientras $(k \leq N)$ haga los pasos 3–9.

Paso 3 Para $i = 1, 2, \ldots, n$ (*Calcule* $\mathbf{r}$.)

determine $r_i = b_i - \sum_{j=1}^{n} a_{ij} x_j$.

(*Realice los cálculos en aritmética de doble precisión* .)

Paso 4 Resuelva el sistema lineal $A\mathbf{y} = \mathbf{r}$ mediante eliminación gaussiana en el mismo orden que en el paso 0.

Paso 5 Para $i = 1, \ldots, n$ determine $xx_i = x_i + y_i$.

Paso 6 Si $k = 1$ entonces determine $COND = \dfrac{\|\mathbf{y}\|_\infty}{\|\mathbf{xx}\|_\infty} 10^t$.

Paso 7 Si $\|\mathbf{x} - \mathbf{xx}\|_\infty < TOL$ entonces SALIDA($\mathbf{xx}$);
SALIDA ($COND$);
(*El procedimiento fue exitoso*.)
PARE.

Paso 8 Determine $k = k + 1$.

Paso 9 Para $i = 1, \ldots, n$ determine $x_i = xx_i$.

Paso 10 SALIDA (‘Número máximo de interacciones extedidas’);
SALIDA ($COND$);
(*El procedimiento no fue exitoso*.)
PARE. ▪

Si se utiliza aritmética de t dígitos, un procedimiento recomendado para interrumpir el proceso en el paso 7 es iterar hasta que $|y_i^{(k)}| \leq 10^{-t}$, para cada $i = 1, 2, \ldots, n$.

Ilustración En nuestra ilustración previa encontramos la aproximación para el sistema lineal

$$\begin{bmatrix} 3.3330 & 15920 & -10.333 \\ 2.2220 & 16.710 & 9.6120 \\ 1.5611 & 5.1791 & 1.6852 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 15913 \\ 28.544 \\ 8.4254 \end{bmatrix}$$

usando la aritmética de cinco dígitos y eliminación gaussiana es

$$\tilde{\mathbf{x}}^{(1)} = (1.2001, 0.99991, 0.92538)^t$$

y la solución para $A\mathbf{y} = \mathbf{r}^{(1)}$ es

$$\tilde{\mathbf{y}}^{(1)} = (-0.20008, 8.9987 \times 10^{-5}, 0.074607)^t.$$

Por el paso 5 en este algoritmo,

$$\tilde{\mathbf{x}}^{(2)} = \tilde{\mathbf{x}}^{(1)} + \tilde{\mathbf{y}}^{(1)} = (1.0000, 1.0000, 0.99999)^t,$$

y el error real en esta aproximación es

$$\|\mathbf{x} - \tilde{\mathbf{x}}^{(2)}\|_\infty = 1 \times 10^{-5}.$$

Usando la técnica para interrumpir el algoritmo, calculamos $\mathbf{r}^{(2)} = \mathbf{b} - A\tilde{\mathbf{x}}^{(2)}$ y resolvemos el sistema $A\mathbf{y}^{(2)} = \mathbf{r}^{(2)}$, que nos da

$$\tilde{\mathbf{y}}^{(2)} = (1.5002 \times 10^{-9}, 2.0951 \times 10^{-10}, 1.0000 \times 10^{-5})^t.$$

Puesto que $\|\tilde{\mathbf{y}}^{(2)}\|_\infty \leq 10^{-5}$, concluimos que

$$\tilde{\mathbf{x}}^{(3)} = \tilde{\mathbf{x}}^{(2)} + \tilde{\mathbf{y}}^{(2)} = (1.0000, 1.0000, 1.0000)^t$$

es suficientemente preciso, lo cual es, sin duda alguna, correcto. ■

A lo largo de esta sección, se ha supuesto que en el sistema lineal $A\mathbf{x} = \mathbf{b}$, A y $\mathbf{b}$ se pueden representar de forma exacta. Siendo realistas, las entradas a_{ij} y b_j, se alterarían o perturbarían por una cantidad de δa_{ij} y δb_j, que causaría que el sistema lineal

$$(A + \delta A)\mathbf{x} = \mathbf{b} + \delta\mathbf{b}$$

se resolviera en lugar de $A\mathbf{x} = \mathbf{b}$. Normalmente, si $\|\delta A\|$ y $\|\delta\mathbf{b}\|$ son pequeñas (en el orden de 10^{-t}), la aritmética de t dígitos produce una solución $\tilde{\mathbf{x}}$ para la que $\|\mathbf{x} - \tilde{\mathbf{x}}\|$ es correspondientemente pequeña. Sin embargo, en caso de sistemas mal condicionados, hemos observado que incluso si A y $\mathbf{b}$ se representan de manera exacta, los errores de redondeo pueden causar que $\|\mathbf{x} - \tilde{\mathbf{x}}\|$ sea grande. El siguiente teorema relaciona las perturbaciones del sistema lineal para el número de condición de una matriz. La prueba de este resultado se puede encontrar en [Or2], p. 33.

Teorema 7.29 Suponga que A es no singular y

$$\|\delta A\| < \frac{1}{\|A^{-1}\|}.$$

La solución $\tilde{\mathbf{x}}$ para $(A + \delta A)\tilde{\mathbf{x}} = \mathbf{b} + \delta\mathbf{b}$ aproxima la solución $\mathbf{x}$ de $A\mathbf{x} = \mathbf{b}$ con el cálculo de error

$$\frac{\|\mathbf{x} - \tilde{\mathbf{x}}\|}{\|\mathbf{x}\|} \leq \frac{K(A)\|A\|}{\|A\| - K(A)\|\delta A\|}\left(\frac{\|\delta\mathbf{b}\|}{\|\mathbf{b}\|} + \frac{\|\delta A\|}{\|A\|}\right). \tag{7.25}$$

■

El cálculo en la desigualdad (7.25) establece que si la matriz A está bien condicionada (es decir, $K(A)$ no es demasiado grande), entonces los cambios pequeños en A y $\mathbf{b}$ producen cambios proporcionalmente pequeños en la solución de $\mathbf{x}$. Si, por otro lado, A está mal condicionada, entonces los cambios pequeños en A y $\mathbf{b}$ pueden producir cambios grandes en $\mathbf{x}$.

El teorema es independiente del procedimiento numérico particular que se usó para resolver $A\mathbf{x} = \mathbf{b}$. Se puede mostrar, mediante un análisis de error regresivo (consulte [Wil1] o [Wil2]), que si la eliminación gaussiana con pivoteo se usa para resolver $A\mathbf{x} = \mathbf{b}$ con aritmética de t dígitos, la solución numérica $\tilde{\mathbf{x}}$ es la solución real de un sistema lineal,

James Hardy Wilkinson (1919–1986) es mejor conocido por su amplio trabajo sobre métodos numéricos para resolver ecuaciones lineales y problemas de eigenvalor. También desarrolló la técnica de análisis de error regresivo.

$$(A + \delta A)\tilde{\mathbf{x}} = \mathbf{b}, \quad \text{donde } \|\delta A\|_\infty \leq f(n)10^{1-t} \max_{i,j,k} |a_{ij}^{(k)}|,$$

para alguna función $f(n)$. En la práctica, Wilkinson descubrió que $f(n) \approx n$ y, en el peor de los casos, que $f(n) \leq 1.01(n^3 + 3n^2)$.

La sección Conjunto de ejercicios 7.5 está disponible en línea. Encuentre la ruta de acceso en las páginas preliminares.

7.6 El método de gradiente conjugado

El método de gradiente conjugado de Hestenes y Stiefel [HS] se desarrolló originalmente como un método directo diseñado para resolver un sistema lineal $n \times n$ definido positivo. Como método directo, en general es inferior a la eliminación gaussiana con pivoteo. Ambos métodos requieren n pasos para determinar una solución y los pasos para el método de gradiente conjugado son más costosos computacionalmente que los de la eliminación gaussiana.

Magnus Hestenes (1906–1991) y Eduard Steifel (1907–1998) publicaron el artículo original sobre el método de gradiente conjugado en 1952 mientras trabajaban en el Instituto para Análisis Numérico en el campus de la UCLA.

Sin embargo, el método de gradiente conjugado es útil cuando se usa como método de aproximación iterativa para resolver los sistemas dispersos grandes con entradas diferentes a cero que se presentan en patrones predecibles. Con frecuencia estos problemas surgen en la solución de problemas de valores en la frontera. Cuando la matriz ha sido precondicionada para realizar cálculos más eficientes, sólo se obtienen buenos resultados en aproximadamente $\sqrt{n}$ iteraciones. Empleado de esta manera, este método es preferible sobre la eliminación gaussiana y los métodos iterativos analizados previamente.

A lo largo de esta sección, suponemos que la matriz A es definida positiva. Usaremos la notación del *producto interno*

$$\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^t \mathbf{y}, \tag{7.26}$$

donde $\mathbf{x}$ y $\mathbf{y}$ son vectores n dimensionales. También necesitaremos algunos resultados estándar adicionales a partir del álgebra lineal. Una revisión de este material se encuentra en la sección 9.1.

El siguiente resultado sigue fácilmente las propiedades de transposiciones (consulte el ejercicio 14).

Teorema 7.30 Para cualquier vector $\mathbf{x}$, $\mathbf{y}$ y $\mathbf{z}$ y cualquier número real α, tenemos

a) $\langle \mathbf{x}, \mathbf{y} \rangle = \langle \mathbf{y}, \mathbf{x} \rangle$; **b)** $\langle \alpha\mathbf{x}, \mathbf{y} \rangle = \langle \mathbf{x}, \alpha\mathbf{y} \rangle = \alpha\langle \mathbf{x}, \mathbf{y} \rangle$;

c) $\langle \mathbf{x} + \mathbf{z}, \mathbf{y} \rangle = \langle \mathbf{x}, \mathbf{y} \rangle + \langle \mathbf{z}, \mathbf{y} \rangle$; **d)** $\langle \mathbf{x}, \mathbf{x} \rangle \geq 0$;

e) $\langle \mathbf{x}, \mathbf{x} \rangle = 0$ si y sólo si $\mathbf{x} = \mathbf{0}$. ■

Cuando A es definida positiva, $\langle \mathbf{x}, A\mathbf{x} \rangle = \mathbf{x}^t A\mathbf{x} > 0$ a menos que $\mathbf{x} = 0$. Además, puesto que A es simétrica, tenemos $\mathbf{x}^t A\mathbf{y} = \mathbf{x}^t A^t\mathbf{y} = (A\mathbf{x})^t\mathbf{y}$, por lo que, además de los resultados en el teorema 7.30, tenemos, para cada $\mathbf{x}$ y $\mathbf{y}$

$$\langle \mathbf{x}, A\mathbf{y} \rangle = (A\mathbf{x})^t\mathbf{y} = \mathbf{x}^t A^t\mathbf{y} = \mathbf{x}^t A\mathbf{y} = \langle A\mathbf{x}, \mathbf{y} \rangle. \tag{7.27}$$

El siguiente resultado es una herramienta básica en el desarrollo del método de gradiente conjugado.

Teorema 7.31 El vector $\mathbf{x}^*$ es una solución para el sistema lineal definido positivo $A\mathbf{x} = \mathbf{b}$ si y sólo si $\mathbf{x}^*$ produce el valor mínimo de

$$g(\mathbf{x}) = \langle \mathbf{x}, A\mathbf{x} \rangle - 2\langle \mathbf{x}, \mathbf{b} \rangle.$$

Demostración Sean $\mathbf{x}$ y $\mathbf{v} \neq \mathbf{0}$ vectores fijos y t una variable de número real. Tenemos

$$\begin{aligned} g(\mathbf{x} + t\mathbf{v}) &= \langle \mathbf{x} + t\mathbf{v}, A\mathbf{x} + tA\mathbf{v} \rangle - 2\langle \mathbf{x} + t\mathbf{v}, \mathbf{b} \rangle \\ &= \langle \mathbf{x}, A\mathbf{x} \rangle + t\langle \mathbf{v}, A\mathbf{x} \rangle + t\langle \mathbf{x}, A\mathbf{v} \rangle + t^2\langle \mathbf{v}, A\mathbf{v} \rangle - 2\langle \mathbf{x}, \mathbf{b} \rangle - 2t\langle \mathbf{v}, \mathbf{b} \rangle \\ &= \langle \mathbf{x}, A\mathbf{x} \rangle - 2\langle \mathbf{x}, \mathbf{b} \rangle + 2t\langle \mathbf{v}, A\mathbf{x} \rangle - 2t\langle \mathbf{v}, \mathbf{b} \rangle + t^2\langle \mathbf{v}, A\mathbf{v} \rangle, \end{aligned}$$

por lo que

$$g(\mathbf{x} + t\mathbf{v}) = g(\mathbf{x}) - 2t\langle \mathbf{v}, \mathbf{b} - A\mathbf{x} \rangle + t^2\langle \mathbf{v}, A\mathbf{v} \rangle. \tag{7.28}$$

Con $\mathbf{x}$ y $\mathbf{v}$ fijos podemos definir la función cuadrática h en t mediante

$$h(t) = g(\mathbf{x} + t\mathbf{v}).$$

Entonces h tiene un valor mínimo cuando $h'(t) = 0$ porque su coeficiente t^2, $\langle \mathbf{v}, A\mathbf{v} \rangle$, es positivo. Puesto que

$$h'(t) = -2\langle \mathbf{v}, \mathbf{b} - A\mathbf{x} \rangle + 2t\langle \mathbf{v}, A\mathbf{v} \rangle,$$

el mínimo se presenta cuando

$$\hat{t} = \frac{\langle \mathbf{v}, \mathbf{b} - A\mathbf{x} \rangle}{\langle \mathbf{v}, A\mathbf{v} \rangle},$$

y, a partir de la ecuación (7.28)

$$\begin{aligned} h(\hat{t}) &= g(\mathbf{x} + \hat{t}\mathbf{v}) \\ &= g(\mathbf{x}) - 2\hat{t}\langle \mathbf{v}, \mathbf{b} - A\mathbf{x} \rangle + \hat{t}^2\langle \mathbf{v}, A\mathbf{v} \rangle \\ &= g(\mathbf{x}) - 2\frac{\langle \mathbf{v}, \mathbf{b} - A\mathbf{x} \rangle}{\langle \mathbf{v}, A\mathbf{v} \rangle}\langle \mathbf{v}, \mathbf{b} - A\mathbf{x} \rangle + \left(\frac{\langle \mathbf{v}, \mathbf{b} - A\mathbf{x} \rangle}{\langle \mathbf{v}, A\mathbf{v} \rangle} \right)^2 \langle \mathbf{v}, A\mathbf{v} \rangle \\ &= g(\mathbf{x}) - \frac{\langle \mathbf{v}, \mathbf{b} - A\mathbf{x} \rangle^2}{\langle \mathbf{v}, A\mathbf{v} \rangle}. \end{aligned}$$

Así, para cualquier vector $\mathbf{v} \neq \mathbf{0}$, tenemos $g(\mathbf{x} + \hat{t}\mathbf{v}) < g(\mathbf{x})$ a menos que $\langle \mathbf{v}, \mathbf{b} - A\mathbf{x} \rangle = 0$, en cuyo caso $g(\mathbf{x}) = g(\mathbf{x} + \hat{t}\mathbf{v})$. Éste es el resultado básico que necesitamos para probar el teorema 7.31.

Suponga que $\mathbf{x}^*$ satisface $A\mathbf{x}^* = \mathbf{b}$. Entonces $\langle \mathbf{v}, \mathbf{b} - A\mathbf{x}^* \rangle = 0$ para cualquier vector $\mathbf{v}$ y $g(\mathbf{x})$ no se puede hacer más pequeño que $g(\mathbf{x}^*)$. Por lo tanto, $\mathbf{x}^*$ minimiza g.

Por otro lado, suponga que $\mathbf{x}^*$ es un vector que minimiza g. Entonces, para cualquier vector $\mathbf{v}$, tenemos $g(\mathbf{x}^* + \hat{t}\mathbf{v}) \geq g(\mathbf{x}^*)$. Por lo tanto, $\langle \mathbf{v}, \mathbf{b} - A\mathbf{x}^* \rangle = 0$. Esto implica que $\mathbf{b} - A\mathbf{x}^* = \mathbf{0}$ y, por consiguiente, que $A\mathbf{x}^* = \mathbf{b}$. ■

Para comenzar el método gradiente conjugado seleccionamos $\mathbf{x}$, una solución aproximada para $A\mathbf{x}^* = \mathbf{b}$ y $\mathbf{v} \neq \mathbf{0}$, lo cual nos da una *dirección de búsqueda* para alejarnos de $\mathbf{x}$ con el fin de mejorar la aproximación. Sea $\mathbf{r} = \mathbf{b} - A\mathbf{x}$ el vector residual relacionado con $\mathbf{x}$ y

$$t = \frac{\langle \mathbf{v}, \mathbf{b} - A\mathbf{x} \rangle}{\langle \mathbf{v}, A\mathbf{v} \rangle} = \frac{\langle \mathbf{v}, \mathbf{r} \rangle}{\langle \mathbf{v}, A\mathbf{v} \rangle}.$$

Si $\mathbf{r} \neq \mathbf{0}$ y $\mathbf{v}$ y $\mathbf{r}$ no son ortogonales, entonces $\mathbf{x} + t\mathbf{v}$ da un valor más pequeño para g que $g(\mathbf{x})$ y es presumiblemente más cercano a $\mathbf{x}^*$ que $\mathbf{x}$. Esto sugiere el siguiente método.

Si $\mathbf{x}^{(0)}$ es una aproximación inicial para $\mathbf{x}^*$ y si $\mathbf{v}^{(1)} \neq \mathbf{0}$ es una dirección de búsqueda inicial. Para $k = 1, 2, 3, \ldots,$ calculamos

$$t_k = \frac{\langle \mathbf{v}^{(k)}, \mathbf{b} - A\mathbf{x}^{(k-1)} \rangle}{\langle \mathbf{v}^{(k)}, A\mathbf{v}^{(k)} \rangle}$$
$$\mathbf{x}^{(k)} = \mathbf{x}^{(k-1)} + t_k \mathbf{v}^{(k)}$$

y seleccionamos una dirección de búsqueda nueva $\mathbf{v}^{(k+1)}$. El objetivo es realizar esta selección de tal forma que la sucesión de aproximaciones $\{\mathbf{x}^{(k)}\}$ converja rápidamente a $\mathbf{x}^*$.

Para seleccionar las direcciones de búsqueda, observamos g como una función de los componentes de $\mathbf{x} = (x_1, x_2, \ldots, x_n)^t$. Por lo tanto,

$$g(x_1, x_2, \ldots, x_n) = \langle \mathbf{x}, A\mathbf{x} \rangle - 2\langle \mathbf{x}, \mathbf{b} \rangle = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} x_i x_j - 2 \sum_{i=1}^{n} x_i b_i.$$

Al tomar derivadas parciales respecto a las variables componentes x_k obtenemos

$$\frac{\partial g}{\partial x_k}(\mathbf{x}) = 2 \sum_{i=1}^{n} a_{ki} x_i - 2b_k,$$

que es el k-ésimo componente del vector $2(A\mathbf{x} - \mathbf{b})$. Por lo tanto, el gradiente de g es

$$\nabla g(\mathbf{x}) = \left(\frac{\partial g}{\partial x_1}(\mathbf{x}), \frac{\partial g}{\partial x_2}(\mathbf{x}), \ldots, \frac{\partial g}{\partial x_n}(\mathbf{x}) \right)^t = 2(A\mathbf{x} - \mathbf{b}) = -2\mathbf{r},$$

donde el vector $\mathbf{r}$ es el vector residual para $\mathbf{x}$.

Del cálculo multivariable, sabemos que la dirección de mayor decrecimiento en el valor de $g(\mathbf{x})$ es la dirección dada por $-\nabla g(\mathbf{x})$, es decir, en la dirección del residuo $\mathbf{r}$. El método que selecciona

$$\mathbf{v}^{(k+1)} = \mathbf{r}^{(k)} = \mathbf{b} - A\mathbf{x}^{(k)}$$

recibe el nombre de *método de descenso rápido.* A pesar de que observaremos en la sección 10.4 que este método tiene mérito para problemas de sistemas no lineales y de optimización, no se usa para sistemas lineales debido a la lenta convergencia.

Un enfoque alternativo utiliza un conjunto de vectores de dirección diferentes a cero $\{\mathbf{v}^{(1)}, \ldots, \mathbf{v}^{(n)}\}$ que satisface

$$\langle \mathbf{v}^{(i)}, A\mathbf{v}^{(j)} \rangle = 0, \quad \text{si} \quad i \neq j.$$

Esto se llama **condición de ortogonalidad de A** y se dice que el conjunto de los vectores $\{\mathbf{v}^{(1)}, \ldots, \mathbf{v}^{(n)}\}$ es **ortogonal a A**. No es difícil mostrar que un conjunto de vectores ortogonales a A, asociados con la matriz definida positiva A, es linealmente independiente. (Consulte el ejercicio 15.) Este conjunto de direcciones de búsqueda da

$$t_k = \frac{\langle \mathbf{v}^{(k)}, \mathbf{b} - A\mathbf{x}^{(k-1)} \rangle}{\langle \mathbf{v}^{(k)}, A\mathbf{v}^{(k)} \rangle} = \frac{\langle \mathbf{v}^{(k)}, \mathbf{r}^{(k-1)} \rangle}{\langle \mathbf{v}^{(k)}, A\mathbf{v}^{(k)} \rangle}$$

y $\mathbf{x}^{(k)} = \mathbf{x}^{(k-1)} + t_k \mathbf{v}^{(k)}$.

El siguiente teorema muestra que esta selección de direcciones de búsqueda provee convergencia en la mayor parte de los n pasos, por lo que como método directo produce la solución exacta, al suponer que la aritmética es exacta.

Teorema 7.32 Sea $\{\mathbf{v}^{(1)}, \ldots, \mathbf{v}^{(n)}\}$ un conjunto de vectores diferentes a cero ortogonal a A relacionados con la matriz definida positiva A y sea $\mathbf{x}^{(0)}$ arbitrario. Defina

$$t_k = \frac{\langle \mathbf{v}^{(k)}, \mathbf{b} - A\mathbf{x}^{(k-1)} \rangle}{\langle \mathbf{v}^{(k)}, A\mathbf{v}^{(k)} \rangle} \quad \text{y} \quad \mathbf{x}^{(k)} = \mathbf{x}^{(k-1)} + t_k \mathbf{v}^{(k)},$$

para $k = 1, 2, \ldots, n$. Entonces, asumiendo la aritmética exacta, $A\mathbf{x}^{(n)} = \mathbf{b}$.

Demostración Puesto que, para cada $k = 1, 2, \ldots, n$, $\mathbf{x}^{(k)} = \mathbf{x}^{(k-1)} + t_k \mathbf{v}^{(k)}$, tenemos

$$\begin{aligned} A\mathbf{x}^{(n)} &= A\mathbf{x}^{(n-1)} + t_n A\mathbf{v}^{(n)} \\ &= (A\mathbf{x}^{(n-2)} + t_{n-1} A\mathbf{v}^{(n-1)}) + t_n A\mathbf{v}^{(n)} \\ &\vdots \\ &= A\mathbf{x}^{(0)} + t_1 A\mathbf{v}^{(1)} + t_2 A\mathbf{v}^{(2)} + \cdots + t_n A\mathbf{v}^{(n)}. \end{aligned}$$

Al restar $\mathbf{b}$ de este resultado obtenemos

$$A\mathbf{x}^{(n)} - \mathbf{b} = A\mathbf{x}^{(0)} - \mathbf{b} + t_1 A\mathbf{v}^{(1)} + t_2 A\mathbf{v}^{(2)} + \cdots + t_n A\mathbf{v}^{(n)}.$$

Ahora tomamos el producto interno a ambos lados con el vector $\mathbf{v}^{(k)}$ y utilizamos las propiedades de los productos internos y el hecho de que A es simétrica para obtener

$$\begin{aligned} \langle A\mathbf{x}^{(n)} - \mathbf{b}, \mathbf{v}^{(k)} \rangle &= \langle A\mathbf{x}^{(0)} - \mathbf{b}, \mathbf{v}^{(k)} \rangle + t_1 \langle A\mathbf{v}^{(1)}, \mathbf{v}^{(k)} \rangle + \cdots + t_n \langle A\mathbf{v}^{(n)}, \mathbf{v}^{(k)} \rangle \\ &= \langle A\mathbf{x}^{(0)} - \mathbf{b}, \mathbf{v}^{(k)} \rangle + t_1 \langle \mathbf{v}^{(1)}, A\mathbf{v}^{(k)} \rangle + \cdots + t_n \langle \mathbf{v}^{(n)}, A\mathbf{v}^{(k)} \rangle. \end{aligned}$$

La propiedad de ortogonalidad de A provee, para cada k,

$$\langle A\mathbf{x}^{(n)} - \mathbf{b}, \mathbf{v}^{(k)} \rangle = \langle A\mathbf{x}^{(0)} - \mathbf{b}, \mathbf{v}^{(k)} \rangle + t_k \langle \mathbf{v}^{(k)}, A\mathbf{v}^{(k)} \rangle. \tag{7.29}$$

Sin embargo, $t_k \langle \mathbf{v}^{(k)}, A\mathbf{v}^{(k)} \rangle = \langle \mathbf{v}^{(k)}, \mathbf{b} - A\mathbf{x}^{(k-1)} \rangle$, luego

$$\begin{aligned} t_k \langle \mathbf{v}^{(k)}, A\mathbf{v}^{(k)} \rangle &= \langle \mathbf{v}^{(k)}, \mathbf{b} - A\mathbf{x}^{(0)} + A\mathbf{x}^{(0)} - A\mathbf{x}^{(1)} + \cdots - A\mathbf{x}^{(k-2)} + A\mathbf{x}^{(k-2)} - A\mathbf{x}^{(k-1)} \rangle \\ &= \langle \mathbf{v}^{(k)}, \mathbf{b} - A\mathbf{x}^{(0)} \rangle + \langle \mathbf{v}^{(k)}, A\mathbf{x}^{(0)} - A\mathbf{x}^{(1)} \rangle + \cdots + \langle \mathbf{v}^{(k)}, A\mathbf{x}^{(k-2)} - A\mathbf{x}^{(k-1)} \rangle. \end{aligned}$$

Pero para cualquier i,

$$\mathbf{x}^{(i)} = \mathbf{x}^{(i-1)} + t_i \mathbf{v}^{(i)} \quad \text{y} \quad A\mathbf{x}^{(i)} = A\mathbf{x}^{(i-1)} + t_i A\mathbf{v}^{(i)},$$

por lo que

$$A\mathbf{x}^{(i-1)} - A\mathbf{x}^{(i)} = -t_i A\mathbf{v}^{(i)}.$$

Por lo tanto,

$$t_k \langle \mathbf{v}^{(k)}, A\mathbf{v}^{(k)} \rangle = \langle \mathbf{v}^{(k)}, \mathbf{b} - A\mathbf{x}^{(0)} \rangle - t_1 \langle \mathbf{v}^{(k)}, A\mathbf{v}^{(1)} \rangle - \cdots - t_{k-1} \langle \mathbf{v}^{(k)}, A\mathbf{v}^{(k-1)} \rangle.$$

Puesto que la ortogonalidad de A $\langle \mathbf{v}^{(k)}, A\mathbf{v}^{(i)} \rangle = 0$, para cada $i \neq k$, entonces

$$\langle \mathbf{v}^{(k)}, A\mathbf{v}^{(k)} \rangle t_k = \langle \mathbf{v}^{(k)}, \mathbf{b} - A\mathbf{x}^{(0)} \rangle.$$

A partir de la ecuación (7.29),

$$\begin{aligned}\langle A\mathbf{x}^{(n)} - \mathbf{b}, \mathbf{v}^{(k)}\rangle &= \langle A\mathbf{x}^{(0)} - \mathbf{b}, \mathbf{v}^{(k)}\rangle + \langle \mathbf{v}^{(k)}, \mathbf{b} - A\mathbf{x}^{(0)}\rangle \\ &= \langle A\mathbf{x}^{(0)} - \mathbf{b}, \mathbf{v}^{(k)}\rangle + \langle \mathbf{b} - A\mathbf{x}^{(0)}, \mathbf{v}^{(k)}\rangle \\ &= \langle A\mathbf{x}^{(0)} - \mathbf{b}, \mathbf{v}^{(k)}\rangle - \langle A\mathbf{x}^{(0)} - \mathbf{b}, \mathbf{v}^{(k)}\rangle = 0.\end{aligned}$$

Por lo tanto, el vector $A\mathbf{x}^{(n)} - \mathbf{b}$ es ortogonal al conjunto de vectores ortogonal de A $\{\mathbf{v}^{(1)}, \ldots, \mathbf{v}^{(n)}\}$. A partir de esto, sigue (consulte el ejercicio 15) que $A\mathbf{x}^{(n)} - \mathbf{b} = \mathbf{0}$, por lo que $A\mathbf{x}^{(n)} = \mathbf{b}$. ■

Ejemplo 1 El sistema lineal

$$\begin{aligned} 4x_1 + 3x_2 \qquad &= 24, \\ 3x_1 + 4x_2 - \quad x_3 &= 30, \\ -\quad x_2 + 4x_3 &= -24, \end{aligned}$$

tiene la solución exacta $\mathbf{x}^* = (3, 4, -5)^t$. Muestre que el procedimiento descrito en el teorema 7.32 con $\mathbf{x}^{(0)} = (0, 0, 0)^t$ produce esta solución exacta después de tres iteraciones.

Solución En el ejemplo 2 de la sección 7.4 establecimos la matriz de coeficientes

$$A = \begin{bmatrix} 4 & 3 & 0 \\ 3 & 4 & -1 \\ 0 & -1 & 4 \end{bmatrix}$$

del sistema es definida positiva. Si $\mathbf{v}^{(1)} = (1, 0, 0)^t$, $\mathbf{v}^{(2)} = (-3/4, 1, 0)^t$, y $\mathbf{v}^{(3)} = (-3/7, 4/7, 1)^t$. Entonces

$$\langle \mathbf{v}^{(1)}, A\mathbf{v}^{(2)}\rangle = \mathbf{v}^{(1)t} A\mathbf{v}^{(2)} = (1, 0, 0) \begin{bmatrix} 4 & 3 & 0 \\ 3 & 4 & -1 \\ 0 & -1 & 4 \end{bmatrix} \begin{bmatrix} -\frac{3}{4} \\ 1 \\ 0 \end{bmatrix} = 0,$$

$$\langle \mathbf{v}^{(1)}, A\mathbf{v}^{(3)}\rangle = (1, 0, 0) \begin{bmatrix} 4 & 3 & 0 \\ 3 & 4 & -1 \\ 0 & -1 & 4 \end{bmatrix} \begin{bmatrix} -\frac{3}{7} \\ \frac{4}{7} \\ 1 \end{bmatrix} = 0,$$

y

$$\langle \mathbf{v}^{(2)}, A\mathbf{v}^{(3)}\rangle = \left(-\frac{3}{4}, 1, 0\right) \begin{bmatrix} 4 & 3 & 0 \\ 3 & 4 & -1 \\ 0 & -1 & 4 \end{bmatrix} \begin{bmatrix} -\frac{3}{7} \\ \frac{4}{7} \\ 1 \end{bmatrix} = 0.$$

Por lo tanto, $\{\mathbf{v}^{(1)}, \mathbf{v}^{(2)}, \mathbf{v}^{(3)}\}$ es un conjunto ortogonal de A.

Al aplicar las iteraciones descritas en el teorema 7.22 para A con $\mathbf{x}^{(0)} = (0, 0, 0)^t$ y $\mathbf{b} = (24, 30, -24)^t$ obtenemos

$$\mathbf{r}^{(0)} = \mathbf{b} - A\mathbf{x}^{(0)} = \mathbf{b} = (24, 30, -24)^t,$$

por lo que

$$\langle \mathbf{v}^{(1)}, \mathbf{r}^{(0)}\rangle = \mathbf{v}^{(1)t}\mathbf{r}^{(0)} = 24, \quad \langle \mathbf{v}^{(1)}, A\mathbf{v}^{(1)}\rangle = 4, \quad \text{y} \quad t_0 = \frac{24}{4} = 6.$$

Por lo tanto,

$$\mathbf{x}^{(1)} = \mathbf{x}^{(0)} + t_0\mathbf{v}^{(1)} = (0, 0, 0)^t + 6(1, 0, 0)^t = (6, 0, 0)^t.$$

Al continuar, tenemos

$$\mathbf{r}^{(1)} = \mathbf{b} - A\mathbf{x}^{(1)} = (0, 12, -24)^t, \quad t_1 = \frac{\langle \mathbf{v}^{(2)}, \mathbf{r}^{(1)} \rangle}{\langle \mathbf{v}^{(2)}, A\mathbf{v}^{(2)} \rangle} = \frac{12}{7/4} = \frac{48}{7},$$

$$\mathbf{x}^{(2)} = \mathbf{x}^{(1)} + t_1\mathbf{v}^{(2)} = (6, 0, 0)^t + \frac{48}{7}\left(-\frac{3}{4}, 1, 0\right)^t = \left(\frac{6}{7}, \frac{48}{7}, 0\right)^t,$$

$$\mathbf{r}^{(2)} = \mathbf{b} - A\mathbf{x}^{(2)} = \left(0, 0, -\frac{120}{7}\right), \quad t_2 = \frac{\langle \mathbf{v}^{(3)}, \mathbf{r}^{(2)} \rangle}{\langle \mathbf{v}^{(3)}, A\mathbf{v}^{(3)} \rangle} = \frac{-120/7}{24/7} = -5,$$

y

$$\mathbf{x}^{(3)} = \mathbf{x}^{(2)} + t_2\mathbf{v}^{(3)} = \left(\frac{6}{7}, \frac{48}{7}, 0\right)^t + (-5)\left(-\frac{3}{7}, \frac{4}{7}, 1\right)^t = (3, 4, -5)^t.$$

Puesto que aplicamos la técnica $n = 3$ veces, ésta debe ser la solución real. ■

Antes de analizar cómo determinar el conjunto ortogonal a A, continuaremos con el desarrollo. El uso de un conjunto $\{\mathbf{v}^{(1)}, \ldots, \mathbf{v}^{(n)}\}$ de vectores de dirección ortogonal a A provee lo que se conoce como método *de dirección conjugada*. El siguiente teorema muestra la ortogonalidad de los vectores residuales $\mathbf{r}^{(k)}$ y de los vectores de dirección $\mathbf{v}^{(j)}$. Una prueba de este resultado mediante inducción matemática se considera en el ejercicio 16.

Teorema 7.33 Los vectores residuales $\mathbf{r}^{(k)}$, donde $k = 1, 2, \ldots, n$, para un método de dirección conjugada, satisfacen las ecuaciones

$$\langle \mathbf{r}^{(k)}, \mathbf{v}^{(j)} \rangle = 0, \quad \text{para cada} \quad j = 1, 2, \ldots, k.$$ ■

El método de gradiente conjugado de Hestenes y Stiefel selecciona las direcciones de búsqueda $\{\mathbf{v}^{(k)}\}$ durante el proceso iterativo de tal forma que los vectores residuales $\{\mathbf{r}^{(k)}\}$ son mutuamente ortogonales. Para construir los vectores de dirección $\{\mathbf{v}^{(1)}, \mathbf{v}^{(2)}, \ldots\}$ y las aproximaciones $\{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots\}$, iniciamos con una aproximación inicial $\mathbf{x}^{(0)}$ y utilizamos la dirección descendente más pronunciada $\mathbf{r}^{(0)} = \mathbf{b} - A\mathbf{x}^{(0)}$ como la primera dirección de búsqueda $\mathbf{v}^{(1)}$.

Suponga que las direcciones conjugadas $\mathbf{v}^{(1)}, \ldots, \mathbf{v}^{(k-1)}$ y las aproximaciones $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(k-1)}$ se han calculado con

$$\mathbf{x}^{(k-1)} = \mathbf{x}^{(k-2)} + t_{k-1}\mathbf{v}^{(k-1)},$$

donde

$$\langle \mathbf{v}^{(i)}, A\mathbf{v}^{(j)} \rangle = 0 \quad \text{y} \quad \langle \mathbf{r}^{(i)}, \mathbf{r}^{(j)} \rangle = 0, \quad \text{para} \quad i \neq j.$$

Si $\mathbf{x}^{(k-1)}$ es la solución para $A\mathbf{x} = \mathbf{b}$, terminamos. De lo contrario, $\mathbf{r}^{(k-1)} = \mathbf{b} - A\mathbf{x}^{(k-1)} \neq \mathbf{0}$, y el teorema 7.33 implica que $\langle \mathbf{r}^{(k-1)}, \mathbf{v}^{(i)} \rangle = 0$, para cada $i = 1, 2, \ldots, k-1$.

Utilizamos $\mathbf{r}^{(k-1)}$ para generar $\mathbf{v}^{(k)}$ al hacer

$$\mathbf{v}^{(k)} = \mathbf{r}^{(k-1)} + s_{k-1}\mathbf{v}^{(k-1)}.$$

Queremos seleccionar s_{k-1} de tal forma que

$$\langle \mathbf{v}^{(k-1)}, A\mathbf{v}^{(k)} \rangle = 0.$$

Puesto que

$$A\mathbf{v}^{(k)} = A\mathbf{r}^{(k-1)} + s_{k-1}A\mathbf{v}^{(k-1)}$$

y

$$\langle \mathbf{v}^{(k-1)}, A\mathbf{v}^{(k)} \rangle = \langle \mathbf{v}^{(k-1)}, A\mathbf{r}^{(k-1)} \rangle + s_{k-1}\langle \mathbf{v}^{(k-1)}, A\mathbf{v}^{(k-1)} \rangle,$$

tendremos $\langle \mathbf{v}^{(k-1)}, A\mathbf{v}^{(k)} \rangle = 0$ cuando

$$s_{k-1} = -\frac{\langle \mathbf{v}^{(k-1)}, A\mathbf{r}^{(k-1)} \rangle}{\langle \mathbf{v}^{(k-1)}, A\mathbf{v}^{(k-1)} \rangle}.$$

También se puede mostrar que con esta selección de s_{k-1}, tenemos $\langle \mathbf{v}^{(k)}, A\mathbf{v}^{(i)} \rangle = 0$, para cada $i = 1, 2, \ldots, k-2$ (consulte [Lu], p. 245). Por lo tanto, $\{\mathbf{v}^{(1)}, \ldots \mathbf{v}^{(k)}\}$ es un conjunto ortogonal a A.

Al haber seleccionado $\mathbf{v}^{(k)}$, calculamos

$$t_k = \frac{\langle \mathbf{v}^{(k)}, \mathbf{r}^{(k-1)} \rangle}{\langle \mathbf{v}^{(k)}, A\mathbf{v}^{(k)} \rangle} = \frac{\langle \mathbf{r}^{(k-1)} + s_{k-1}\mathbf{v}^{(k-1)}, \mathbf{r}^{(k-1)} \rangle}{\langle \mathbf{v}^{(k)}, A\mathbf{v}^{(k)} \rangle}$$
$$= \frac{\langle \mathbf{r}^{(k-1)}, \mathbf{r}^{(k-1)} \rangle}{\langle \mathbf{v}^{(k)}, A\mathbf{v}^{(k)} \rangle} + s_{k-1}\frac{\langle \mathbf{v}^{(k-1)}, \mathbf{r}^{(k-1)} \rangle}{\langle \mathbf{v}^{(k)}, A\mathbf{v}^{(k)} \rangle}.$$

Mediante el teorema 7.33, $\langle \mathbf{v}^{(k-1)}, \mathbf{r}^{(k-1)} \rangle = 0$, por lo que

$$t_k = \frac{\langle \mathbf{r}^{(k-1)}, \mathbf{r}^{(k-1)} \rangle}{\langle \mathbf{v}^{(k)}, A\mathbf{v}^{(k)} \rangle}. \tag{7.30}$$

Por lo tanto,

$$\mathbf{x}^{(k)} = \mathbf{x}^{(k-1)} + t_k\mathbf{v}^{(k)}.$$

Para calcular $\mathbf{r}^{(k)}$, multiplicamos A y restamos $\mathbf{b}$ para obtener

$$A\mathbf{x}^{(k)} - \mathbf{b} = A\mathbf{x}^{(k-1)} - \mathbf{b} + t_k A\mathbf{v}^{(k)}$$

o

$$\mathbf{r}^{(k)} = \mathbf{r}^{(k-1)} - t_k A\mathbf{v}^{(k)}.$$

Esto da

$$\langle \mathbf{r}^{(k)}, \mathbf{r}^{(k)} \rangle = \langle \mathbf{r}^{(k-1)}, \mathbf{r}^{(k)} \rangle - t_k\langle A\mathbf{v}^{(k)}, \mathbf{r}^{(k)} \rangle = -t_k\langle \mathbf{r}^{(k)}, A\mathbf{v}^{(k)} \rangle.$$

Además, a partir de la ecuación (7.30)

$$\langle \mathbf{r}^{(k-1)}, \mathbf{r}^{(k-1)} \rangle = t_k\langle \mathbf{v}^{(k)}, A\mathbf{v}^{(k)} \rangle,$$

por lo que

$$s_k = -\frac{\langle \mathbf{v}^{(k)}, A\mathbf{r}^{(k)} \rangle}{\langle \mathbf{v}^{(k)}, A\mathbf{v}^{(k)} \rangle} = -\frac{\langle \mathbf{r}^{(k)}, A\mathbf{v}^{(k)} \rangle}{\langle \mathbf{v}^{(k)}, A\mathbf{v}^{(k)} \rangle} = \frac{(1/t_k)\langle \mathbf{r}^{(k)}, \mathbf{r}^{(k)} \rangle}{(1/t_k)\langle \mathbf{r}^{(k-1)}, \mathbf{r}^{(k-1)} \rangle} = \frac{\langle \mathbf{r}^{(k)}, \mathbf{r}^{(k)} \rangle}{\langle \mathbf{r}^{(k-1)}, \mathbf{r}^{(k-1)} \rangle}.$$

En resumen, tenemos

$$\mathbf{r}^{(0)} = \mathbf{b} - A\mathbf{x}^{(0)}; \quad \mathbf{v}^{(1)} = \mathbf{r}^{(0)};$$

y para $k = 1, 2, \ldots, n$,

$$t_k = \frac{\langle \mathbf{r}^{(k-1)}, \mathbf{r}^{(k-1)} \rangle}{\langle \mathbf{v}^{(k)}, A\mathbf{v}^{(k)} \rangle}, \quad \mathbf{x}^{(k)} = \mathbf{x}^{(k-1)} + t_k\mathbf{v}^{(k)}, \mathbf{r}^{(k)} = \mathbf{r}^{(k-1)} - t_k A\mathbf{v}^{(k)}, s_k = \frac{\langle \mathbf{r}^{(k)}, \mathbf{r}^{(k)} \rangle}{\langle \mathbf{r}^{(k-1)}, \mathbf{r}^{(k-1)} \rangle},$$

y

$$\mathbf{v}^{(k+1)} = \mathbf{r}^{(k)} + s_k\mathbf{v}^{(k)}. \tag{7.31}$$

Precondicionamiento

En lugar de presentar un algoritmo para el método de gradiente conjugado mediante estas fórmulas ampliamos el método para incluir *precondicionamiento*. Si la matriz A está mal condicionada, el método de gradiente conjugado es altamente susceptible a errores de redondeo. Por lo tanto, a pesar de que se obtendría la respuesta exacta en los n pasos, normalmente, éste no es el caso. Como método directo, el método de gradiente conjugado no es tan bueno como la eliminación gaussiana con pivoteo. El uso principal del método de gradiente conjugado es un método iterativo aplicado a un sistema mejor condicionado. En este caso, con frecuencia se obtiene una solución aproximada aceptable alrededor de $\sqrt{n}$ pasos.

El precondicioamiento reemplaza un sistema determinado por uno que tiene las mismas soluciones, pero con mejores características de convergencia.

Cuando se usa precondicionamiento, el método de gradiente conjugado no se aplica directamente a la matriz A, sino a otra matriz definida positiva que tiene un número de condición más pequeño. Necesitamos hacer esto de tal forma que una vez que se encuentra la solución de este sistema, será fácil obtener la solución para el sistema original. La expectativa es que esto reducirá el error de redondeo al aplicar el método. Para mantener la definición positiva de la matriz resultante, necesitamos multiplicar en cada lado por una matriz no singular. Denotaremos esta matriz mediante C^{-1} y consideraremos

$$\tilde{A} = C^{-1}A(C^{-1})^t,$$

con la esperanza de que $\tilde{A}$ tenga un número de condición menor que A. Para simplificar la notación, utilizamos notación de matriz $C^{-t} \equiv \left(C^{-1}\right)^t$. Más adelante en esta sección, observaremos una forma razonable de seleccionar C, pero primero consideraremos el método de gradiente conjugado aplicado a $\tilde{A}$.

Considere el sistema lineal

$$\tilde{A}\tilde{\mathbf{x}} = \tilde{\mathbf{b}},$$

donde $\tilde{\mathbf{x}} = C^t\mathbf{x}$ y $\tilde{\mathbf{b}} = C^{-1}\mathbf{b}$. Entonces,

$$\tilde{A}\tilde{\mathbf{x}} = (C^{-1}AC^{-t})(C^t\mathbf{x}) = C^{-1}A\mathbf{x}.$$

Por lo tanto, podemos resolver $\tilde{A}\tilde{\mathbf{x}} = \tilde{\mathbf{b}}$ para $\tilde{\mathbf{x}}$ y, después, obtener $\mathbf{x}$ al multiplicar por C^{-t}. Sin embargo, en lugar de reescribir la ecuación (7.31) mediante $\tilde{\mathbf{r}}^{(k)}, \tilde{\mathbf{v}}^{(k)}, \tilde{t}_k, \tilde{\mathbf{x}}^{(k)}$ y $\tilde{s}_k$, incluimos implícitamente la precondición.

Puesto que

$$\tilde{\mathbf{x}}^{(k)} = C^t\mathbf{x}^{(k)},$$

tenemos

$$\tilde{\mathbf{r}}^{(k)} = \tilde{\mathbf{b}} - \tilde{A}\tilde{\mathbf{x}}^{(k)} = C^{-1}\mathbf{b} - (C^{-1}AC^{-t})C^t\mathbf{x}^{(k)} = C^{-1}(\mathbf{b} - A\mathbf{x}^{(k)}) = C^{-1}\mathbf{r}^{(k)}.$$

Si $\tilde{\mathbf{v}}^{(k)} = C^t\mathbf{v}^{(k)}$ y $\mathbf{w}^{(k)} = C^{-1}\mathbf{r}^{(k)}$. Entonces

$$\tilde{s}_k = \frac{\langle \tilde{\mathbf{r}}^{(k)}, \tilde{\mathbf{r}}^{(k)} \rangle}{\langle \tilde{\mathbf{r}}^{(k-1)}, \tilde{\mathbf{r}}^{(k-1)} \rangle} = \frac{\langle C^{-1}\mathbf{r}^{(k)}, C^{-1}\mathbf{r}^{(k)} \rangle}{\langle C^{-1}\mathbf{r}^{(k-1)}, C^{-1}\mathbf{r}^{(k-1)} \rangle},$$

por lo que

$$\tilde{s}_k = \frac{\langle \mathbf{w}^{(k)}, \mathbf{w}^{(k)} \rangle}{\langle \mathbf{w}^{(k-1)}, \mathbf{w}^{(k-1)} \rangle}. \tag{7.32}$$

Por lo tanto,

$$\tilde{t}_k = \frac{\langle \tilde{\mathbf{r}}^{(k-1)}, \tilde{\mathbf{r}}^{(k-1)} \rangle}{\langle \tilde{\mathbf{v}}^{(k)}, \tilde{A}\tilde{\mathbf{v}}^{(k)} \rangle} = \frac{\langle C^{-1}\mathbf{r}^{(k-1)}, C^{-1}\mathbf{r}^{(k-1)} \rangle}{\langle C^t\mathbf{v}^{(k)}, C^{-1}AC^{-t}C^t\mathbf{v}^{(k)} \rangle} = \frac{\langle \mathbf{w}^{(k-1)}, \mathbf{w}^{(k-1)} \rangle}{\langle C^t\mathbf{v}^{(k)}, C^{-1}A\mathbf{v}^{(k)} \rangle}$$

y puesto que

$$\langle C^t\mathbf{v}^{(k)}, C^{-1}A\mathbf{v}^{(k)}\rangle = [C^t\mathbf{v}^{(k)}]^t C^{-1}A\mathbf{v}^{(k)}$$
$$= [\mathbf{v}^{(k)}]^t CC^{-1}A\mathbf{v}^{(k)} = [\mathbf{v}^{(k)}]^t A\mathbf{v}^{(k)} = \langle \mathbf{v}^{(k)}, A\mathbf{v}^{(k)}\rangle,$$

tenemos

$$\tilde{t}_k = \frac{\langle \mathbf{w}^{(k-1)}, \mathbf{w}^{(k-1)}\rangle}{\langle \mathbf{v}^{(k)}, A\mathbf{v}^{(k)}\rangle}. \tag{7.33}$$

Además

$$\tilde{\mathbf{x}}^{(k)} = \tilde{\mathbf{x}}^{(k-1)} + \tilde{t}_k\tilde{\mathbf{v}}^{(k)}, \quad \text{entonces } C^t\mathbf{x}^{(k)} = C^t\mathbf{x}^{(k-1)} + \tilde{t}_k C^t\mathbf{v}^{(k)}$$

y

$$\mathbf{x}^{(k)} = \mathbf{x}^{(k-1)} + \tilde{t}_k\mathbf{v}^{(k)}. \tag{7.34}$$

Al continuar,

$$\tilde{\mathbf{r}}^{(k)} = \tilde{\mathbf{r}}^{(k-1)} - \tilde{t}_k\tilde{A}\tilde{\mathbf{v}}^{(k)},$$

por lo tanto

$$C^{-1}\mathbf{r}^{(k)} = C^{-1}\mathbf{r}^{(k-1)} - \tilde{t}_k C^{-1}AC^{-t}\tilde{v}^{(k)}, \qquad \mathbf{r}^{(k)} = \mathbf{r}^{(k-1)} - \tilde{t}_k AC^{-t}C^t\mathbf{v}^{(k)},$$

y

$$\mathbf{r}^{(k)} = \mathbf{r}^{(k-1)} - \tilde{t}_k A\mathbf{v}^{(k)}. \tag{7.35}$$

Finalmente,

$$\tilde{\mathbf{v}}^{(k+1)} = \tilde{\mathbf{r}}^{(k)} + \tilde{s}_k\tilde{\mathbf{v}}^{(k)} \quad \text{y} \quad C^t\mathbf{v}^{(k+1)} = C^{-1}\mathbf{r}^{(k)} + \tilde{s}_k C^t\mathbf{v}^{(k)},$$

por lo que

$$\mathbf{v}^{(k+1)} = C^{-t}C^{-1}\mathbf{r}^{(k)} + \tilde{s}_k\mathbf{v}^{(k)} = C^{-t}\mathbf{w}^{(k)} + \tilde{s}_k\mathbf{v}^{(k)}. \tag{7.36}$$

El método gradiente conjugado precondicionado está basado en el uso de las ecuaciones (7.32) a (7.36) en orden (7.33), (7.34), (7.35), (7.32) y (7.36). El algoritmo 7.5 implementa este procedimiento.

ALGORITMO **7.5**

Método de gradiente conjugado precondicionado

Para resolver $A\mathbf{x} = \mathbf{b}$ dada la matriz precondicionada C^{-1} y la aproximación inicial $\mathbf{x}^{(0)}$:

ENTRADA el número de ecuaciones y valores desconocidos n; las entradas a_{ij}, $1 \le i, j \le n$ de la matriz A; las entradas b_j, $1 \le j \le n$ del vector $\mathbf{b}$; las entradas γ_{ij}, $1 \le i, j \le n$ de la matriz precondicionada C^{-1}, las entradas x_i, $1 \le i \le n$ de la aproximación inicial $\mathbf{x} = \mathbf{x}^{(0)}$, el número máximo de iteraciones N; la tolerancia TOL.

SALIDA la solución aproximada $x_1, \ldots x_n$ y el residuo $r_1, \ldots r_n$ o un mensaje de que se excedió el número de iteraciones.

Paso 1 Determine $\mathbf{r} = \mathbf{b} - A\mathbf{x}$; (*Calcule* $\mathbf{r}^{(0)}$)
$\mathbf{w} = C^{-1}\mathbf{r}$; (*Nota:* $\mathbf{w} = \mathbf{w}^{(0)}$)
$\mathbf{v} = C^{-t}\mathbf{w}$; (*Nota:* $\mathbf{v} = \mathbf{v}^{(1)}$)
$\alpha = \sum_{j=1}^{n} w_j^2$.

Paso 2 Determine $k = 1$.

Paso 3 Mientras $(k \le N)$ haga los pasos 4–7.

Paso 3 Mientras ($k \leq N$) haga los pasos 4–7.

Paso 4 Si $\|\mathbf{v}\| < TOL$, entonces
SALIDA ('Vector solución'; $x_1, \ldots, x_n$);
SALIDA ('Con residual'; $r_1, \ldots, r_n$);
(*El procedimiento fue exitoso*.)
PARE.

Paso 5 Determine $\mathbf{u} = A\mathbf{v}$; (*Nota:* $\mathbf{u} = A\mathbf{v}^{(k)}$)
$t = \dfrac{\alpha}{\sum_{j=1}^{n} v_j u_j}$; (*Nota:* $t = t_k$)
$\mathbf{x} = \mathbf{x} + t\mathbf{v}$; (*Nota:* $\mathbf{x} = \mathbf{x}^{(k)}$)
$\mathbf{r} = \mathbf{r} - t\mathbf{u}$; (*Nota:* $\mathbf{r} = \mathbf{r}^{(k)}$)
$\mathbf{w} = C^{-1}\mathbf{r}$; (*Nota:* $\mathbf{w} = \mathbf{w}^{(k)}$)
$\beta = \sum_{j=1}^{n} w_j^2$. (*Nota:* $\beta = \langle \mathbf{w}^{(k)}, \mathbf{w}^{(k)} \rangle$)

Paso 6 Si $|\beta| < TOL$ entonces
si $\|\mathbf{r}\| < TOL$ entonces
SALIDA ('Vector solución'; $x_1, \ldots, x_n$);
SALIDA ('con residuo'; $r_1, \ldots, r_n$);
(*El procedimiento fue exitoso*.)
PARE.

Paso 7 Determine $s = \beta/\alpha$; ($s = s_k$)
$\mathbf{v} = C^{-t}\mathbf{w} + s\mathbf{v}$; (*Nota:* $\mathbf{v} = \mathbf{v}^{(k+1)}$)
$\alpha = \beta$; (*Actualice* α.)
$k = k + 1$.

Paso 8 Si ($k > n$) entonces
SALIDA ('Se excedió el máximo número de iteraciones');
(*El procedimiento no fue exitoso*.)
PARE. ■

El siguiente ejemplo ilustra los cálculos para un problema elemental.

Ejemplo 2 El sistema lineal $A\mathbf{x} = \mathbf{b}$ dado por

$$\begin{aligned} 4x_1 + 3x_2 &= 24, \\ 3x_1 + 4x_2 - x_3 &= 30, \\ - x_2 + 4x_3 &= -24 \end{aligned}$$

tiene solución $(3, 4, -5)^t$. Utilice el método de gradiente conjugado con $\mathbf{x}^{(0)} = (0, 0, 0)^t$ y sin precondicionamiento, es decir, con $C = C^{-1} = I$, para aproximar la solución.

Solución La solución se consideró en el ejemplo 2 de la sección 7.4 donde el método SOR se utilizó con un valor casi óptimo de $\omega = 1.25$.

Para el método de gradiente conjugado iniciamos con

$$\mathbf{r}^{(0)} = \mathbf{b} - A\mathbf{x}^{(0)} = \mathbf{b} = (24, 30, -24)^t;$$

$$\mathbf{w} = C^{-1}\mathbf{r}^{(0)} = (24, 30, -24)^t;$$

$$\mathbf{v}^{(1)} = C^{-t}\mathbf{w} = (24, 30, -24)^t;$$

$$\alpha = \langle \mathbf{w}, \mathbf{w} \rangle = 2052.$$

Iniciamos con la primera iteración con $k = 1$. Entonces,

$$\mathbf{u} = A\mathbf{v}^{(1)} = (186.0, 216.0, -126.0)^t;$$

$$t_1 = \frac{\alpha}{\langle \mathbf{v}^{(1)}, \mathbf{u} \rangle} = 0.1469072165;$$

$$\mathbf{x}^{(1)} = \mathbf{x}^{(0)} + t_1\mathbf{v}^{(1)} = (3.525773196, 4.407216495, -3.525773196)^t;$$

$$\mathbf{r}^{(1)} = \mathbf{r}^{(0)} - t_1\mathbf{u} = (-3.32474227, -1.73195876, -5.48969072)^t;$$

$$\mathbf{w} = C^{-1}\mathbf{r}^{(1)} = \mathbf{r}^{(1)};$$

$$\beta = \langle \mathbf{w}, \mathbf{w} \rangle = 44.19029651;$$

$$s_1 = \frac{\beta}{\alpha} = 0.02153523222;$$

$$\mathbf{v}^{(2)} = C^{-t}\mathbf{w} + s_1\mathbf{v}^{(1)} = (-2.807896697, -1.085901793, -6.006536293)^t.$$

Establezca

$$\alpha = \beta = 44.19029651.$$

Para la segunda iteración, tenemos

$$\mathbf{u} = A\mathbf{v}^{(2)} = (-14.48929217, -6.760760967, -22.94024338)^t;$$

$$t_2 = 0.2378157558;$$

$$\mathbf{x}^{(2)} = (2.858011121, 4.148971939, -4.954222164)^t;$$

$$\mathbf{r}^{(2)} = (0.121039698, -0.124143281, -0.034139402)^t;$$

$$\mathbf{w} = C^{-1}\mathbf{r}^{(2)} = \mathbf{r}^{(2)};$$

$$\beta = 0.03122766148;$$

$$s_2 = 0.0007066633163;$$

$$\mathbf{v}^{(3)} = (0.1190554504, -0.1249106480, -0.03838400086)^t.$$

Determina $\alpha = \beta = 0.03122766148$.

La tercera iteración da

$$\mathbf{u} = A\mathbf{v}^{(3)} = (0.1014898976, -0.1040922099, -0.0286253554)^t;$$

$$t_3 = 1.192628008;$$

$$\mathbf{x}^{(3)} = (2.999999998, 4.000000002, -4.999999998)^t;$$

$$\mathbf{r}^{(3)} = (0.36 \times 10^{-8}, 0.39 \times 10^{-8}, -0.141 \times 10^{-8})^t.$$

Puesto que $\mathbf{x}^{(3)}$ es aproximadamente la solución exacta, el error de redondeo no afecta significativamente el resultado. En el ejemplo 2 de la sección 7.4, el método SOR con $\omega = 1.25$ requería 14 iteraciones para una precisión de 10^{-7}. Sin embargo, se debería observar, en este ejemplo que en verdad estamos comparando un método directo para métodos iterativos. ■

El siguiente ejemplo ilustra el efecto del precondicionamiento en una matriz pobremente condicionada. En este ejemplo, usamos $D^{-1/2}$ para representar la matriz diagonal cuyas entradas son los recíprocos de las raíces cuadradas de las entradas de la diagonal de la matriz A de coeficientes. Ésto se utiliza como precondicionador. Puesto que la matriz A es definida positiva, esperamos que los eigevalores de $D^{-1/2}AD^{-t/2}$ estén cerca de 1 con el resultado de que el número de condición de esta matriz sería relativamente pequeño para el número de condición de A.

Ejemplo 3 Encuentre los eigenvalores y número de condición de la matriz

$$A = \begin{bmatrix} 0.2 & 0.1 & 1 & 1 & 0 \\ 0.1 & 4 & -1 & 1 & -1 \\ 1 & -1 & 60 & 0 & -2 \\ 1 & 1 & 0 & 8 & 4 \\ 0 & -1 & -2 & 4 & 700 \end{bmatrix}$$

y compárelos con los eigenvalores y números de condición de la matriz precondicionada $D^{-1/2}AD^{-t/2}$.

Solución Para determinar la matriz precondicionada, primero necesitamos la matriz diagonal, la cual, al ser simétrica también es su transpuesta. Sus entradas diagonales se especifican mediante

$$a1 = \frac{1}{\sqrt{0.2}};\ a2 = \frac{1}{\sqrt{4.0}};\ a3 = \frac{1}{\sqrt{60.0}};\ a4 = \frac{1}{\sqrt{8.0}};\ a5 = \frac{1}{\sqrt{700.0}},$$

y la matriz de precondicionamiento es

$$C^{-1} = \begin{bmatrix} 2.23607 & 0 & 0 & 0 & 0 \\ 0 & .500000 & 0 & 0 & 0 \\ 0 & 0 & .129099 & 0 & 0 \\ 0 & 0 & 0 & .353553 & 0 \\ 0 & 0 & 0 & 0 & 0.0377965 \end{bmatrix}.$$

La matriz precondicionada es

$$\tilde{A} = C^{-1}AC^{-t}$$

$$= \begin{bmatrix} 1.000002 & 0.1118035 & 0.2886744 & 0.7905693 & 0 \\ 0.1118035 & 1 & -0.0645495 & 0.1767765 & -0.0188983 \\ 0.2886744 & -0.0645495 & 0.9999931 & 0 & -0.00975898 \\ 0.7905693 & 0.1767765 & 0 & 0.9999964 & 0.05345219 \\ 0 & -0.0188983 & -0.00975898 & 0.05345219 & 1.000005 \end{bmatrix}.$$

Se encuentra que los eigenvalores de A y $\tilde{A}$ son

Eigenvalores de A 700.031, 60.0284, 0.0570747, 8.33845, 3.74533 y

Eigenvalores de $\tilde{A}$ 1.88052, 0.156370, 0.852686, 1.10159, 1.00884.

Los números de condición de A y $\tilde{A}$ con la norma l_∞ que se encuentran son 13961.7 para A y 16.1155 para $\tilde{A}$. Sin duda alguna, es verdad que en este caso $\tilde{A}$ está mejor condicionada que la matriz original A. ■

Ilustración El sistema $A\mathbf{x} = \mathbf{b}$ con

$$A = \begin{bmatrix} 0.2 & 0.1 & 1 & 1 & 0 \\ 0.1 & 4 & -1 & 1 & -1 \\ 1 & -1 & 60 & 0 & -2 \\ 1 & 1 & 0 & 8 & 4 \\ 0 & -1 & -2 & 4 & 700 \end{bmatrix} \quad \text{y} \quad \mathbf{b} = \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{bmatrix}$$

tiene la solución

$$\mathbf{x}^* = (7.859713071, 0.4229264082, -0.07359223906, -0.5406430164, 0.01062616286)^t.$$

La tabla 7.5 muestra los resultados obtenidos al utilizar los métodos iterativos de Jacobi, Gauss-Siedel y SOR (con $\omega = 1.25$) para el sistema con A con una tolerancia de 0.01, así como aquellos cuando el método de gradiente conjugado se aplica tanto en su forma no precondicionada como mediante la matriz de precondicionamiento descrita en el ejemplo 3. El método de gradiente conjugado no sólo provee las aproximaciones más precisas, sino que también utiliza un número más pequeño de iteraciones. ■

Tabla 7.5

Método	Número de iteraciones	$\mathbf{x}^{(k)}$	$\|\mathbf{x}^* - \mathbf{x}^{(k)}\|_\infty$
Jacobi	49	$(7.86277141, 0.42320802, -0.07348669, -0.53975964, 0.01062847)^t$	0.00305834
Gauss-Seidel	15	$(7.83525748, 0.42257868, -0.07319124, -0.53753055, 0.01060903)^t$	0.02445559
SOR ($\omega = 1.25$)	7	$(7.85152706, 0.42277371, -0.07348303, -0.53978369, 0.01062286)^t$	0.00818607
Gradiente conjugado	5	$(7.85341523, 0.42298677, -0.07347963, -0.53987920, 0.008628916)^t$	0.00629785
Gradiente conjugado (precondicionado)	4	$(7.85968827, 0.42288329, -0.07359878, -0.54063200, 0.01064344)^t$	0.00009312

A menudo, el método de gradiente conjugado precondicionado se utiliza en la solución de los grandes sistemas lineales en los que la matriz está dispersa y es definida positiva. Estos sistemas se deben resolver para soluciones aproximadas para problemas de valores en la frontera de ecuaciones diferenciales ordinarias (secciones 11.3, 11.4 y 11.5). Mientras más grande sea el sistema, más prometedor será el método de gradiente conjugado porque reduce significativamente el número de iteraciones requeridas. En estos sistemas, la matriz precondicionada C es aproximadamente igual a L en la factorización de Cholesky LL^t de A. En general, las entradas pequeñas en A son ignoradas y el método de Cholesky se aplica para obtener lo que recibe el nombre de una factorización LL^t incompleta de A. Por lo tanto, $C^{-t}C^{-1} \approx A^{-1}$, y se obtiene una buena aproximación. Más información sobre el método de gradiente conjugado se puede encontrar en [Kelley].

La sección Conjunto de ejercicios 7.6 está disponible en línea. Encuentre la ruta de acceso en las páginas preliminares.



7.7 Software numérico

Casi todos los paquetes comerciales y de dominio público que contienen métodos iterativos para la solución de un sistema de ecuaciones lineales requieren el uso de un precondicionador con el método. A menudo, la rápida convergencia de los solucionadores se logra al utilizar un precondicionador. Un precondicionador produce un sistema equivalente de ecuaciones que, con suerte, presenta mejores características de convergencia que el sistema original. La Biblioteca IMSL tiene un método de gradiente conjugado precondicionado y la Biblioteca NAG tiene varias subrutinas, que son prefijos, para la solución iterativa de sistemas lineales.

Aleksei Nikolaevich Krylov (1863–1945) trabajó en matemáticas aplicadas, principalmente en las áreas de problemas de valores en la frontera, la aceleración de la convergencia de series de Fourier y varios problemas clásicos relacionados con sistemas mecánicos. Durante principios de la década de 1930 fue el director del Instituto de Física-Matemáticas de la Academia Soviética de Ciencias.

Todas las subrutinas están basadas en subespacios de Krylov. Saad [Sa2] tiene una descripción detallada de los métodos de subespacios Krylov. Los paquetes LINPACK y LAPACK sólo contienen métodos directos para la solución de sistemas lineales; sin embargo, los paquetes contienen muchas subrutinas que se utilizan mediante solucionadores iterativos. Los paquetes de dominio público IML++, ITPACK, SLAP y Templates contienen métodos iterativos. MATLAB contiene varios métodos iterativos que también están basados en subespacios Krylov.

Los conceptos de número de condición y matrices pobremente condicionadas se introdujeron en la sección 7.5. Muchas de las subrutinas para resolver un sistema lineal o para factorizar una matriz en una factorización *LU* incluyendo verificaciones para matrices mal condicionadas, y también proporcionan un cálculo del número de condición. LAPACK tiene numerosas rutinas que incluyen el cálculo de un número de condición, como lo hacen las bibliotecas ISML y NAG.

Las bibliotecas LAPACK, LINPACK, IMSL y NAG tienen subrutinas que mejoran una solución para un sistema lineal que está pobremente condicionado. La subrutina prueba el número de condición y, después, utiliza el refinamiento iterativo para obtener la solución más precisa posible dada la precisión de la computadora.

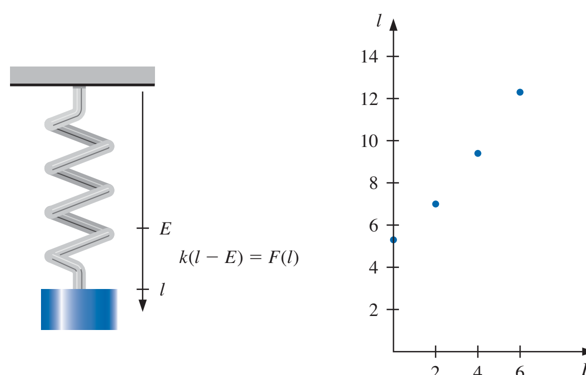
Las secciones Preguntas de análisis, Conceptos clave y Revisión del capítulo están disponibles en línea. Encuentre la ruta de acceso en las páginas preliminares.

CAPÍTULO

8 Teoría de aproximación

Introducción

La ley de Hooke establece que cuando se aplica una fuerza a un resorte construido con material uniforme, la longitud del resorte es una función lineal de esa fuerza. Podemos escribir la función lineal como $F(l) = k(l - E)$, donde $F(l)$ representa la fuerza requerida para estirar el resorte l unidades, la constante E representa la longitud del resorte sin fuerza aplicada y la constante k es la constante del resorte.



Suponga que queremos determinar la constante para un resorte que tiene una longitud inicial de 5.3 pulgadas. Aplicamos fuerzas de 2, 4 y 6 libras al resorte y encontramos que su longitud aumenta a 7.0, 9.4 y 12.3 pulgadas, respectivamente. Una revisión rápida muestra que los puntos (0, 5.3), (2, 7.0), (4, 9.4) y (6, 12.3) no se encuentran completamente en línea recta. Aunque podríamos usar un par aleatorio de estos puntos de datos para aproximar la constante del resorte, parecería más razonable encontrar la recta que *mejor* aproxima a todos los puntos de datos para determinar la constante. En este capítulo se considerará este tipo de aproximación y es posible encontrar esta aplicación de resorte en el ejercicio 7 de la sección 8.1.

La teoría de la aproximación implica dos tipos generales de problemas. Uno surge cuando una función se define de manera explícita, pero nos gustaría encontrar un tipo de función "más simple", como un polinomio, para los valores aproximados de una función determinada. El otro problema se preocupa por ajustar funciones a un dato establecido y encontrar la "mejor" función de cierta clase para representar los datos.

Ambos problemas se han analizado en el capítulo 3. El enésimo polinomio de Taylor alrededor del número x_0 es una excelente aproximación para una función $f\,(n + 1)$ veces diferenciable en una vecindad de x_0. Los polinomios de interpolación de Lagrange o, de modo más general, osculantes, se analizaron como polinomios de interpolación y para ajustar ciertos datos. Los splines cúbicos también se analizaron en el capítulo 3. En este capítulo, se consideran las limitaciones para estas técnicas y se analizan otras vías de enfoque.

8.1 Aproximación por mínimos cuadrados discretos

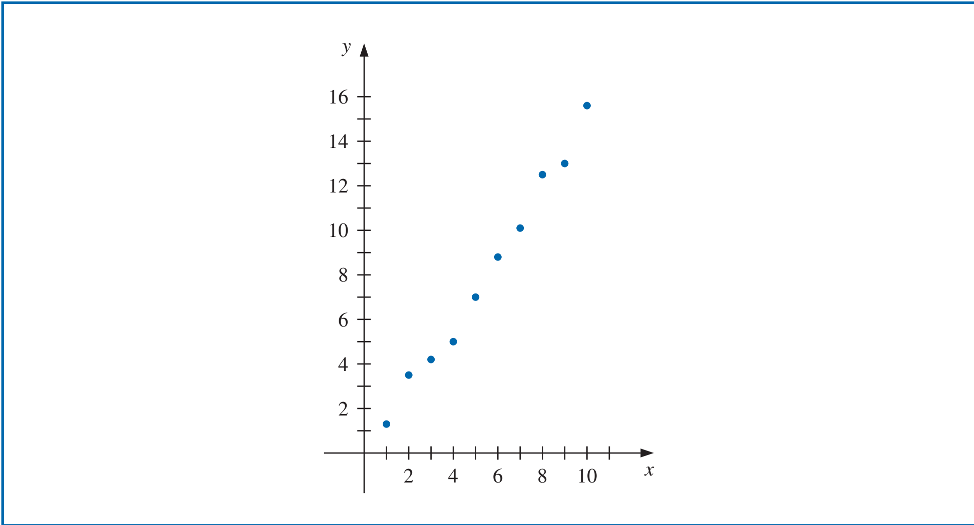
Considere el problema de calcular los valores de una función en puntos no tabulados, dados los datos experimentales en la tabla 8.1.

Tabla 8.1

x_i	y_i	x_i	y_i
1	1.3	6	8.8
2	3.5	7	10.1
3	4.2	8	12.5
4	5.0	9	13.0
5	7.0	10	15.6

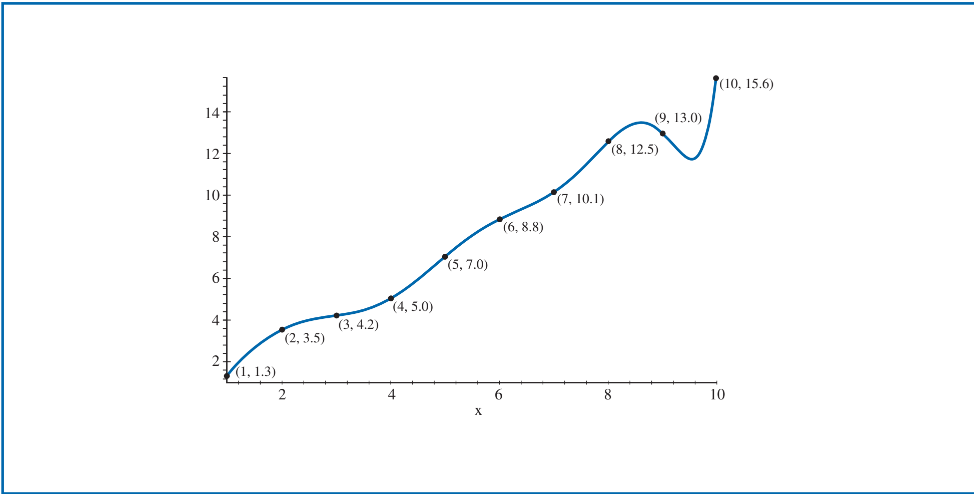
La figura 8.1 muestra una gráfica de los valores de la tabla 8.1. A partir de esta gráfica, parece que la relación real entre x y y es lineal. La razón probable para que ninguna línea se ajuste con precisión a los datos son los errores en estos últimos. Por lo que es poco razonable solicitar que la función de aproximación concuerde exactamente con los datos. De hecho, dicha función introduciría oscilaciones que no estaban presentes originalmente. Por ejemplo, la gráfica del polinomio de interpolación de noveno grado que se muestra de modo libre para los datos en la tabla 8.1 se muestra en la figura 8.2.

Figura 8.1



La gráfica obtenida (con puntos de datos adicionales) se muestra en la figura 8.2.

Figura 8.2



Este polinomio es claramente una predicción de la información entre una serie de puntos de datos. Un mejor enfoque sería encontrar la recta que se aproxima "mejor" (en cierto sentido), incluso si no concuerda precisamente con los datos en ningún punto.

Sea que $a_1x_i + a_0$ denota el i-ésimo valor en la recta de aproximación y que y_i es el i-ésimo valor de y dado. Suponemos que las variables independientes, las x_i, son exactas; son las variables dependientes, las y_i, de las que sospechamos. Esto es una suposición razonable en muchas situaciones experimentales.

El problema de encontrar la ecuación de la mejor aproximación lineal en el sentido absoluto requiere encontrar los valores a_0 y a_1 para minimizar

$$E_\infty(a_0, a_1) = \max_{1\leq i\leq 10}\{|y_i - (a_1x_i + a_0)|\}.$$

Normalmente esto recibe el nombre de problema **minimáx** y no es posible manejarlo con técnicas fundamentales.

Otro enfoque para determinar la mejor aproximación lineal implica encontrar los valores de a_0 y a_1 para minimizar

$$E_1(a_0, a_1) = \sum_{i=1}^{10} |y_i - (a_1x_i + a_0)|.$$

Esta cantidad recibe el nombre de **desviación absoluta**. Para minimizar una función de dos variables, necesitamos igualar sus derivadas parciales a cero y resolver simultáneamente las ecuaciones resultantes. En el caso de la desviación absoluta, necesitamos encontrar a_0 y a_1 con

$$0 = \frac{\partial}{\partial a_0}\sum_{i=1}^{10} |y_i - (a_1x_i + a_0)| \quad \text{y} \quad 0 = \frac{\partial}{\partial a_1}\sum_{i=1}^{10} |y_i - (a_1x_i + a_0)|.$$

El problema es que la función valor absoluto no es diferenciable en cero y podríamos no encontrar soluciones para este par de ecuaciones.

Mínimos cuadrados lineales

El enfoque de **mínimos cuadrados** para este problema implica determinar la mejor línea de aproximación cuando el error relacionado es la suma de los cuadrados de las diferencias entre los valores y en la línea de aproximación y los valores y proporcionados. Por lo tanto, deben encontrarse las constantes a_0 y a_1 que minimizan el error de mínimos cuadrados:

$$E_2(a_0, a_1) = \sum_{i=1}^{10} [y_i - (a_1x_i + a_0)]^2 .$$

El método de mínimos cuadrados es el procedimiento más conveniente para determinar mejores aproximaciones lineales, pero también hay consideraciones teóricas importantes que lo favorecen. En general, mientras el enfoque minimáx asigna demasiado peso a un bit de datos con un gran error, el método de desviación absoluta no da suficiente peso a un punto que está fuera de la línea con la aproximación. El enfoque de mínimos cuadrados asigna considerablemente más peso en un punto que está fuera de la línea que al resto de los datos, pero no permitirá que el punto domine por completo la aproximación. Una razón adicional para considerar el enfoque de mínimos cuadrados implica el estudio de la distribución estadística del error (consulte [Lar], p. 463–481).

El problema general de ajustar la mejor línea de mínimos cuadrados para una recopilación de datos $\{(x_i, y_i)\}_{i=1}^m$ implica minimizar el error total,

$$E \equiv E_2(a_0, a_1) = \sum_{i=1}^{m} [y_i - (a_1x_i + a_0)]^2 ,$$

respecto a los parámetros a_0 y a_1. Para que se presente un mínimo, necesitamos que

$$\frac{\partial E}{\partial a_0} = 0 \quad \text{y} \quad \frac{\partial E}{\partial a_1} = 0,$$

es decir,

$$0 = \frac{\partial}{\partial a_0} \sum_{i=1}^{m} [(y_i - (a_1 x_i - a_0)]^2 = 2 \sum_{i=1}^{m} (y_i - a_1 x_i - a_0)(-1)$$

y

$$0 = \frac{\partial}{\partial a_1} \sum_{i=1}^{m} [y_i - (a_1 x_i + a_0)]^2 = 2 \sum_{i=1}^{m} (y_i - a_1 x_i - a_0)(-x_i).$$

La palabra "normal" como aquí se usa implica la idea de "perpendicular". Las ecuaciones normales se obtienen encontrando direcciones perpendiculares para una superficie multidimensional.

Estas ecuaciones se simplifican en las **ecuaciones normales**:

$$a_0 \cdot m + a_1 \sum_{i=1}^{m} x_i = \sum_{i=1}^{m} y_i \quad \text{y} \quad a_0 \sum_{i=1}^{m} x_i + a_1 \sum_{i=1}^{m} x_i^2 = \sum_{i=1}^{m} x_i y_i.$$

La solución para este sistema de ecuaciones es

$$a_0 = \frac{\sum_{i=1}^{m} x_i^2 \sum_{i=1}^{m} y_i - \sum_{i=1}^{m} x_i y_i \sum_{i=1}^{m} x_i}{m\left(\sum_{i=1}^{m} x_i^2\right) - \left(\sum_{i=1}^{m} x_i\right)^2} \tag{8.1}$$

y

$$a_1 = \frac{m \sum_{i=1}^{m} x_i y_i - \sum_{i=1}^{m} x_i \sum_{i=1}^{m} y_i}{m\left(\sum_{i=1}^{m} x_i^2\right) - \left(\sum_{i=1}^{m} x_i\right)^2}. \tag{8.2}$$

Ejemplo 1 Encuentre la línea de mínimos cuadrados que se aproxima a los datos en la tabla 8.1.

Solución Primero ampliamos la tabla para incluir x_i^2 y $x_i y_i$ y sumamos las columnas. Esto se muestra en la tabla 8.2.

Tabla 8.2

x_i	y_i	x_i^2	$x_i y_i$	$P(x_i) = 1.538x_i - 0.360$
1	1.3	1	1.3	1.18
2	3.5	4	7.0	2.72
3	4.2	9	12.6	4.25
4	5.0	16	20.0	5.79
5	7.0	25	35.0	7.33
6	8.8	36	52.8	8.87
7	10.1	49	70.7	10.41
8	12.5	64	100.0	11.94
9	13.0	81	117.0	13.48
10	15.6	100	156.0	15.02
55	81.0	385	572.4	$E = \sum_{i=1}^{10} (y_i - P(x_i))^2 \approx 2.34$

Las ecuaciones normales (8.1) y (8.2) implican que

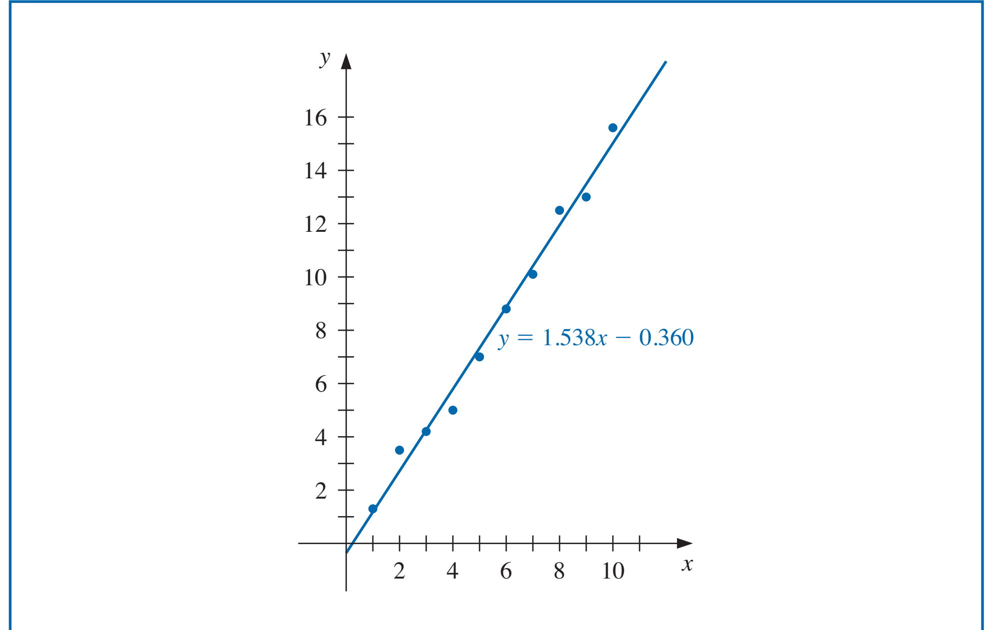
$$a_0 = \frac{385(81) - 55(572.4)}{10(385) - (55)^2} = -0.360$$

y

$$a_1 = \frac{10(572.4) - 55(81)}{10(385) - (55)^2} = 1.538,$$

por lo que $P(x) = 1.538x - 0.360$. La gráfica de esta recta y los puntos de datos se muestran en la figura 8.3. Los valores aproximados obtenidos por la técnica de mínimos cuadrados en los puntos de datos están en la tabla 8.2. ■

Figura 8.3



Mínimos cuadrados polinomiales

El problema general de aproximar un conjunto de datos $\{(x_i, y_i) \mid i = 1, 2, \ldots, m\}$, con un polinomio algebraico

$$P_n(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0,$$

de grado $n < m - 1$, por medio del procedimiento de mínimos cuadrados se maneja de forma similar. Seleccionamos las constantes $a_0, a_1, \ldots, a_n$ para minimizar el error de mínimos cuadrados $E = E_2(a_0, a_1, \ldots, a_n)$, donde

$$\begin{aligned} E &= \sum_{i=1}^{m} (y_i - P_n(x_i))^2 \\ &= \sum_{i=1}^{m} y_i^2 - 2\sum_{i=1}^{m} P_n(x_i)y_i + \sum_{i=1}^{m} (P_n(x_i))^2 \end{aligned}$$

$$= \sum_{i=1}^{m} y_i^2 - 2\sum_{i=1}^{m}\left(\sum_{j=0}^{n} a_j x_i^j\right) y_i + \sum_{i=1}^{m}\left(\sum_{j=0}^{n} a_j x_i^j\right)^2$$

$$= \sum_{i=1}^{m} y_i^2 - 2\sum_{j=0}^{n} a_j\left(\sum_{i=1}^{m} y_i x_i^j\right) + \sum_{j=0}^{n}\sum_{k=0}^{n} a_j a_k\left(\sum_{i=1}^{m} x_i^{j+k}\right).$$

Como en el caso lineal, para minimizar E es necesario que $\partial E/\partial a_j = 0$, para cada $j = 0, 1, \ldots, n$. Por lo tanto, para cada j, debemos tener

$$0 = \frac{\partial E}{\partial a_j} = -2\sum_{i=1}^{m} y_i x_i^j + 2\sum_{k=0}^{n} a_k \sum_{i=1}^{m} x_i^{j+k}.$$

Esto nos da $n + 1$ **ecuaciones normales** en las $n + 1$ incógnitas a_j. Éstas son

$$\sum_{k=0}^{n} a_k \sum_{i=1}^{m} x_i^{j+k} = \sum_{i=1}^{m} y_i x_i^j, \quad \text{para cada } j = 0, 1, \ldots, n. \tag{8.3}$$

Es útil escribir las ecuaciones de acuerdo con lo siguiente:

$$a_0\sum_{i=1}^{m} x_i^0 + a_1\sum_{i=1}^{m} x_i^1 + a_2\sum_{i=1}^{m} x_i^2 + \cdots + a_n\sum_{i=1}^{m} x_i^n = \sum_{i=1}^{m} y_i x_i^0,$$

$$a_0\sum_{i=1}^{m} x_i^1 + a_1\sum_{i=1}^{m} x_i^2 + a_2\sum_{i=1}^{m} x_i^3 + \cdots + a_n\sum_{i=1}^{m} x_i^{n+1} = \sum_{i=1}^{m} y_i x_i^1,$$

$$\vdots$$

$$a_0\sum_{i=1}^{m} x_i^n + a_1\sum_{i=1}^{m} x_i^{n+1} + a_2\sum_{i=1}^{m} x_i^{n+2} + \cdots + a_n\sum_{i=1}^{m} x_i^{2n} = \sum_{i=1}^{m} y_i x_i^n.$$

Estas *ecuaciones normales* tienen una única solución siempre y cuando las x_i sean distintas (consulte el ejercicio 14).

Ejemplo 2 Ajuste los datos en la tabla 8.3 con el polinomio de mínimos cuadrados discretos de grado máximo 2.

Tabla 8.3

i	x_i	y_i
1	0	1.0000
2	0.25	1.2840
3	0.50	1.6487
4	0.75	2.1170
5	1.00	2.7183

Solución Para este problema, $n = 2$, $m = 5$, y las tres ecuaciones normales son

$$5a_0 + 2.5a_1 + 1.875a_2 = 8.7680,$$
$$2.5a_0 + 1.875a_1 + 1.5625a_2 = 5.4514, \text{ y}$$
$$1.875a_0 + 1.5625a_1 + 1.3828a_2 = 4.4015.$$

Al resolver las ecuaciones obtenemos

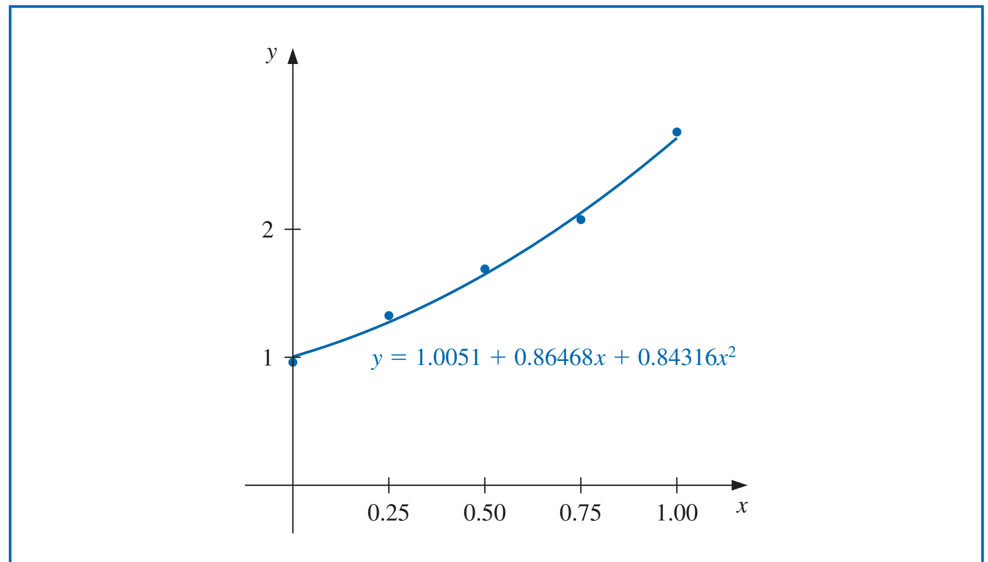
$$a_0 = 1.005075519, \quad a_1 = 0.8646758482, \text{ y } \quad a_2 = 0.8431641518.$$

Por lo tanto, el polinomio de mínimos cuadrados de grado 2 que se ajusta a los datos de la tabla 8.3 es

$$P_2(x) = 1.0051 + 0.86468x + 0.84316x^2,$$

cuya gráfica se muestra en la figura 8.4. En los valores determinados de x_i, tenemos las aproximaciones mostradas en la tabla 8.4.

Figura 8.4



El error total,

$$E = \sum_{i=1}^{5} (y_i - P(x_i))^2 = 2.74 \times 10^{-4},$$

es el mínimo que se puede obtener a través de un polinomio de grado máximo 2. ■

Tabla 8.4

i	1	2	3	4	5
x_i	0	0.25	0.50	0.75	1.00
y_i	1.0000	1.2840	1.6487	2.1170	2.7183
$P(x_i)$	1.0051	1.2740	1.6482	2.1279	2.7129
$y_i - P(x_i)$	−0.0051	0.0100	0.0004	−0.0109	0.0054

Algunas veces es adecuado asumir que los datos están exponencialmente relacionados. Esto requiere que la función de aproximación sea de la forma

$$y = be^{ax} \tag{8.4}$$

o

$$y = bx^a, \tag{8.5}$$

para algunas constantes a y b. La dificultad de aplicar el procedimiento de mínimos cuadrados en una situación de este tipo proviene de intentar minimizar

$$E = \sum_{i=1}^{m} (y_i - be^{ax_i})^2, \quad \text{en el caso de la ecuación (8.4)}$$

o

$$E = \sum_{i=1}^{m} (y_i - bx_i^a)^2, \qquad \text{en el caso de la ecuación (8.5).}$$

Las ecuaciones normales relacionadas con estos procedimientos se obtienen ya sea a partir de

$$0 = \frac{\partial E}{\partial b} = 2\sum_{i=1}^{m} (y_i - be^{ax_i})(-e^{ax_i})$$

y

$$0 = \frac{\partial E}{\partial a} = 2\sum_{i=1}^{m} (y_i - be^{ax_i})(-bx_i e^{ax_i}), \quad \text{en el caso de la ecuación (8.4),}$$

o

$$0 = \frac{\partial E}{\partial b} = 2\sum_{i=1}^{m} (y_i - bx_i^a)(-x_i^a)$$

y

$$0 = \frac{\partial E}{\partial a} = 2\sum_{i=1}^{m} (y_i - bx_i^a)(-b(\ln x_i)x_i^a), \quad \text{en el caso de la ecuación (8.5).}$$

En general, no se puede encontrar una solución exacta para estos sistemas en a y b.

El método que se utiliza normalmente cuando se sospecha que los datos están exponencialmente relacionados es considerar el logaritmo de la ecuación de aproximación:

$$\ln y = \ln b + ax, \quad \text{en el caso de la ecuación (8.4),}$$

y

$$\ln y = \ln b + a \ln x, \quad \text{en el caso de la ecuación (8.5).}$$

En cualquier caso, ahora aparece un problema lineal y las soluciones para $\ln b$ y a se pueden obtener al modificar adecuadamente las ecuaciones normales (8.1) y (8.2).

Sin embargo, la obtenida de esta forma *no* es la aproximación por mínimos cuadrados para el problema original, y esta aproximación puede, en algunos casos, diferir significativamente de la aproximación de mínimos cuadrados para el problema original. La aplicación en el ejercicio 13 describe este problema. Esta aplicación se reconsiderará en el ejercicio 9 en la sección 10.3, donde la solución exacta para el problema exponencial de mínimos cuadrados se aproxima con métodos adecuados para resolver sistemas de ecuaciones no lineales.

Ilustración Considere el conjunto de datos en las primeras tres columnas de la tabla 8.5.

Tabla 8.5

i	x_i	y_i	$\ln y_i$	x_i^2	$x_i \ln y_i$
1	1.00	5.10	1.629	1.0000	1.629
2	1.25	5.79	1.756	1.5625	2.195
3	1.50	6.53	1.876	2.2500	2.814
4	1.75	7.45	2.008	3.0625	3.514
5	2.00	8.46	2.135	4.0000	4.270
	7.50		9.404	11.875	14.422

Si x_i se grafica con ln y_i, los datos parecen tener una relación lineal, por lo que es razonable suponer una aproximación de la forma

$$y = be^{ax}, \qquad \text{lo cual implica que} \qquad \ln y = \ln b + ax.$$

Al expandir la tabla y suponer que las columnas apropiadas dan los datos restantes en la tabla 8.5.

Usando las ecuaciones normales (8.1) y (8.2),

$$a = \frac{(5)(14.422) - (7.5)(9.404)}{(5)(11.875) - (7.5)^2} = 0.5056$$

y

$$\ln b = \frac{(11.875)(9.404) - (14.422)(7.5)}{(5)(11.875) - (7.5)^2} = 1.122.$$

Con $\ln b = 1.222$, tenemos $b = e^{1.122} = 3.071$, y la aproximación asume la forma

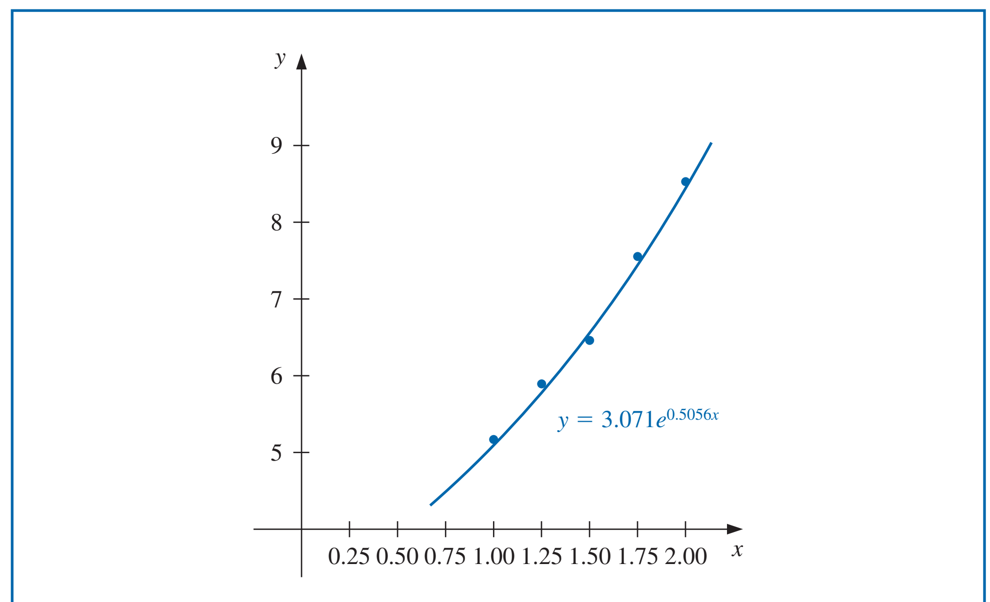
$$y = 3.071e^{0.5056x_i}.$$

En los puntos de datos, estos valores se dan en la tabla 8.6 (consulte la figura 8.5.) ■

Tabla 8.6

i	x_i	y_i	$3.071e^{0.5056x_i}$	$\lvert y_i - 3.071e^{0.5056x_i}\rvert$
1	1.00	5.10	5.09	0.01
2	1.25	5.79	5.78	0.01
3	1.50	6.53	6.56	0.03
4	1.75	7.45	7.44	0.01
5	2.00	8.46	8.44	0.02

Figura 8.5



La sección Conjunto de ejercicios 8.1 está disponible en línea. Encuentre la ruta de acceso en las páginas preliminares.

8.2 Polinomios ortogonales y aproximación por mínimos cuadrados

La sección previa consideraba el problema de la aproximación por mínimos cuadrados para ajustarse a un conjunto de datos. El otro problema de aproximación mencionado en la introducción aborda la aproximación de funciones.

Suponga que $f \in C[a, b]$ y que se requiere un polinomio $P_n(x)$ de grado a lo sumo n para minimizar el error

$$\int_a^b [f(x) - P_n(x)]^2 \, dx.$$

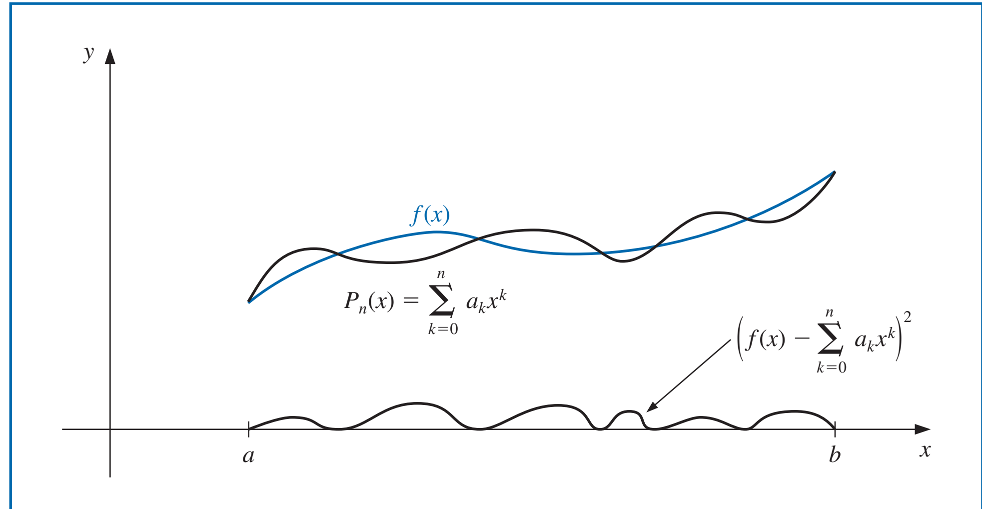
Para determinar un polinomio de aproximación por mínimos cuadrados, es decir, un polinomio para minimizar esta aproximación, sea

$$P_n(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0 = \sum_{k=0}^{n} a_k x^k$$

y defina, como se muestra en la figura 8.6,

$$E \equiv E_2(a_0, a_1, \ldots, a_n) = \int_a^b \left(f(x) - \sum_{k=0}^{n} a_k x^k \right)^2 dx.$$

Figura 8.6



El problema es encontrar los coeficientes reales $a_0, a_1, \ldots, a_n$ que minimizarán E. Una condición necesaria para los números $a_0, a_1, \ldots, a_n$ para minimizar E es que

$$\frac{\partial E}{\partial a_j} = 0, \quad \text{para cada } j = 0, 1, \ldots, n.$$

Puesto que

$$E = \int_a^b [f(x)]^2 \, dx - 2\sum_{k=0}^{n} a_k \int_a^b x^k f(x) \, dx + \int_a^b \left(\sum_{k=0}^{n} a_k x^k \right)^2 dx,$$

tenemos

$$\frac{\partial E}{\partial a_j} = -2\int_a^b x^j f(x) \, dx + 2\sum_{k=0}^{n} a_k \int_a^b x^{j+k} \, dx.$$

Por lo tanto, encontramos $P_n(x)$, las **ecuaciones normales** lineales $(n + 1)$

$$\sum_{k=0}^{n} a_k \int_a^b x^{j+k}\, dx = \int_a^b x^j f(x)\, dx, \text{ para cada } \quad j = 0, 1, \ldots, n, \tag{8.6}$$

se deben resolver para las $(n + 1)$ incógnitas a_j. Las ecuaciones normales siempre tienen una única solución siempre y cuando $f \in C[a, b]$. (Consulte el ejercicio 15.)

Ejemplo 1 Encuentre el polinomio de aproximación de grado 2 por mínimos cuadrados para la función $f(x) = \operatorname{sen} \pi x$ en el intervalo [0, 1].

Solución Las ecuaciones normales para $P_2(x) = a_2x^2 + a_1x + a_0$ son

$$a_0 \int_0^1 1\, dx + a_1 \int_0^1 x\, dx + a_2 \int_0^1 x^2\, dx = \int_0^1 \operatorname{sen} \pi x\, dx,$$

$$a_0 \int_0^1 x\, dx + a_1 \int_0^1 x^2\, dx + a_2 \int_0^1 x^3\, dx = \int_0^1 x \operatorname{sen} \pi x\, dx, \quad \text{y}$$

$$a_0 \int_0^1 x^2\, dx + a_1 \int_0^1 x^3\, dx + a_2 \int_0^1 x^4\, dx = \int_0^1 x^2 \operatorname{sen} \pi x\, dx.$$

Realizando la integración obtenemos

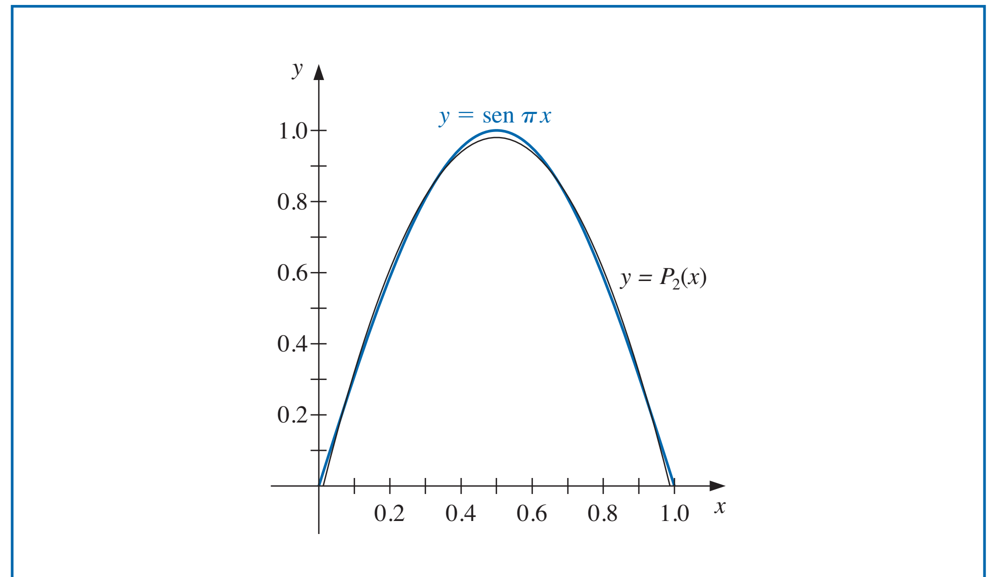
$$a_0 + \frac{1}{2}a_1 + \frac{1}{3}a_2 = \frac{2}{\pi}, \quad \frac{1}{2}a_0 + \frac{1}{3}a_1 + \frac{1}{4}a_2 = \frac{1}{\pi}, \quad \text{y} \quad \frac{1}{3}a_0 + \frac{1}{4}a_1 + \frac{1}{5}a_2 = \frac{\pi^2 - 4}{\pi^3}.$$

Estas tres ecuaciones en tres incógnitas se pueden resolver para obtener

$$a_0 = \frac{12\pi^2 - 120}{\pi^3} \approx -0.050465 \quad \text{y} \quad a_1 = -a_2 = \frac{720 - 60\pi^2}{\pi^3} \approx 4.12251.$$

Por consiguiente, la aproximación del polinomio de grado 2 por mínimos cuadrados para $f(x) = \operatorname{sen} \pi x$ en [0, 1] es $P_2(x) = -4.12251x^2 + 4.12251x - 0.050465$. (Consulte la figura 8.7.) ■

Figura 8.7



David Hilbert (1862–1943) fue un matemático dominante a finales del siglo xx. Se le recuerda mejor por impartir una charla en el Congreso Internacional de Matemáticos, en París, en 1900, donde planteó 23 problemas que había pensado que sería importante que los matemáticos del siglo siguiente resolvieran.

El ejemplo 1 ilustra una dificultad al obtener una aproximación polinomial de mínimos cuadrados. Se debe resolver un sistema lineal $(n + 1) \times (n + 1)$ para las incógnitas $a_0, \ldots, a_n$ y los coeficientes en el sistema lineal son de la forma

$$\int_a^b x^{j+k}\, dx = \frac{b^{j+k+1} - a^{j+k+1}}{j+k+1},$$

un sistema lineal que no tiene una solución numérica fácil de calcular. La matriz en el sistema lineal es conocida como **matriz de Hilbert**, un ejemplo clásico para demostrar dificultades de error de redondeo (consulte el ejercicio 9 de la sección 7.5).

Otra desventaja es similar a la situación que se presentó cuando los polinomios de Lagrange se presentaron por primera vez en la sección 3.1. Los cálculos realizados para obtener el mejor polinomio de enésimo grado $P_n(x)$, no reducen la cantidad de trabajo requerido para obtener $P_{n+1}(x)$, el polinomio del siguiente grado superior.

Funciones linealmente independientes

Ahora consideraremos una técnica diferente para obtener aproximaciones de mínimos cuadrados. Ésta resulta ser eficiente desde el punto de vista informático y una vez que se conoce $P_n(x)$ es sencillo determinar $P_{n+1}(x)$ Para facilitar el análisis, necesitamos algunos conceptos nuevos.

Definición 8.1 Se dice que el conjunto de funciones $\{\phi_0, \ldots, \phi_n\}$ es **linealmente independiente** en $[a, b]$ si,

$$c_0\phi_0(x) + c_1\phi_1(x) + \cdots + c_n\phi_n(x) = 0, \qquad \text{para todas las } x \in [a, b],$$

entonces $c_0 = c_1 = \cdots = c_n = 0$. De lo contrario, se dice que el conjunto de funciones es **linealmente dependiente**. ■

Teorema 8.2 Suponga que, para cada $j = 0, 1, \ldots, n$, $\phi_j(x)$ es un polinomio de grado j. Entonces el conjunto $\{\phi_0, \ldots, \phi_n\}$ es linealmente independiente en cualquier intervalo $[a, b]$.

Demostración Sean $c_0, \ldots, c_n$ números reales para los que

$$P(x) = c_0\phi_0(x) + c_1\phi_1(x) + \cdots + c_n\phi_n(x) = 0, \quad \text{para todos } x \in [a, b].$$

El polinomio $P(x)$ se anula en $[a, b]$, por lo que debe ser el polinomio cero y los coeficientes de todas las potencias de x son cero. En particular, el coeficiente de x^n es cero. Pero $c_n\phi_n(x)$ es el único término en $P(x)$ que contiene x^n, por lo que debemos tener $c_n = 0$. Por lo tanto

$$P(x) = \sum_{j=0}^{n-1} c_j\phi_j(x).$$

En esta representación de $P(x)$, el único término que contiene una potencia de x^{n-1} es $c_{n-1}\phi_{n-1}(x)$, por lo que este término también debe ser cero y

$$P(x) = \sum_{j=0}^{n-2} c_j\phi_j(x).$$

De la misma forma, las constantes restantes $c_{n-2}, c_{n-3}, \ldots, c_1, c_0$ son cero, lo cual implica que $\{\phi_0, \phi_1, \ldots, \phi_n\}$ es linealmente independiente en $[a, b]$. ■

Ejemplo 2 Si $\phi_0(x) = 2$, $\phi_1(x) = x-3$, y $\phi_2(x) = x^2+2x+7$, y $Q(x) = a_0+a_1x+a_2x^2$. Muestre que existen constantes c_0, c_1, y c_2 tales que $Q(x) = c_0\phi_0(x) + c_1\phi_1(x) + c_2\phi_2(x)$.

Solución Por el teorema 8.2 $\{\phi_0, \phi_1, \phi_2\}$ es linealmente independiente en cualquier intervalo $[a, b]$. Primero observe que

$$1 = \frac{1}{2}\phi_0(x), \quad x = \phi_1(x) + 3 = \phi_1(x) + \frac{3}{2}\phi_0(x)$$

y que

$$x^2 = \phi_2(x) - 2x - 7 = \phi_2(x) - 2\left[\phi_1(x) + \frac{3}{2}\phi_0(x)\right] - 7\left[\frac{1}{2}\phi_0(x)\right]$$
$$= \phi_2(x) - 2\phi_1(x) - \frac{13}{2}\phi_0(x).$$

Por lo tanto,

$$Q(x) = a_0\left[\frac{1}{2}\phi_0(x)\right] + a_1\left[\phi_1(x) + \frac{3}{2}\phi_0(x)\right] + a_2\left[\phi_2(x) - 2\phi_1(x) - \frac{13}{2}\phi_0(x)\right]$$
$$= \left(\frac{1}{2}a_0 + \frac{3}{2}a_1 - \frac{13}{2}a_2\right)\phi_0(x) + [a_1 - 2a_2]\,\phi_1(x) + a_2\phi_2(x).$$ ■

La situación que se ilustra en el ejemplo 2 se mantiene en una configuración mucho más general. Si $\prod_n$ denota el **conjunto de todos los polinomios de grado a lo sumo *n***. El siguiente resultado se utiliza ampliamente en muchas aplicaciones de álgebra lineal. Su demostración se considera en el ejercicio 13.

Teorema 8.3 Suponga que $\{\phi_0(x), \phi_1(x), \ldots, \phi_n(x)\}$ es un conjunto de polinomios linealmente independientes en $\prod_n$. Entonces, un polinomio en $\prod_n$ se puede escribir de manera única como una combinación lineal de $\phi_0(x), \phi_1(x), \ldots, \phi_n(x)$. ■

Funciones ortogonales

Analizar la aproximación general de una función requiere la introducción de las nociones de función de peso y ortogonalidad.

Definición 8.4 Una función integrable w recibe el nombre de **función de peso** en el intervalo I si $w(x) \geq 0$, para todas las x en I, pero $w(x) \not\equiv 0$ en cualquier subintervalo de I. ■

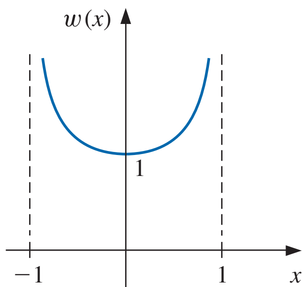
El objetivo de una función de peso es asignar varios grados de importancia a las aproximaciones en ciertas partes del intervalo. Por ejemplo, la función de peso

$$w(x) = \frac{1}{\sqrt{1 - x^2}}$$

asigna menos énfasis cerca del centro del intervalo $(-1, 1)$ y más énfasis cuando $|x|$ está cerca de 1 (consulte la figura 8.8). Esta función de peso se usa en la siguiente sección.

Suponga que $\{\phi_0, \phi_1, \ldots, \phi_n\}$ es un conjunto de funciones linealmente independiente en $[a, b]$ y w es una función de peso para $[a, b]$. Dada $f \in C[a, b]$, buscamos una combinación lineal

Figura 8.8



$$P(x) = \sum_{k=0}^{n} a_k\phi_k(x)$$

para minimizar el error

$$E = E(a_0, \ldots, a_n) = \int_a^b w(x)\left[f(x) - \sum_{k=0}^{n} a_k\phi_k(x)\right]^2 dx.$$

Este problema reduce la situación considerada al inicio de esta sección en el caso especial cuando $w(x) \equiv 1$ y $\phi_k(x) = x^k$, para cada $k = 0, 1, \ldots, n$.

Las ecuaciones normales relacionadas con este problema se derivan del hecho de que para cada $j = 0, 1, \ldots, n$,

$$0 = \frac{\partial E}{\partial a_j} = 2\int_a^b w(x)\left[f(x) - \sum_{k=0}^{n} a_k\phi_k(x)\right]\phi_j(x)\,dx.$$

El sistema de ecuaciones normales se puede escribir como

$$\int_a^b w(x)f(x)\phi_j(x)\,dx = \sum_{k=0}^{n} a_k \int_a^b w(x)\phi_k(x)\phi_j(x)\,dx, \quad \text{para } j = 0, 1, \ldots, n.$$

Si las funciones $\phi_0, \phi_1, \ldots, \phi_n$ se pueden seleccionar de tal forma que

$$\int_a^b w(x)\phi_k(x)\phi_j(x)\,dx = \begin{cases} 0, & \text{cuando } j \neq k, \\ \alpha_j > 0, & \text{cuando } j = k, \end{cases} \tag{8.7}$$

entonces las ecuaciones normales se reducirán a

$$\int_a^b w(x)f(x)\phi_j(x)\,dx = a_j\int_a^b w(x)[\phi_j(x)]^2\,dx = a_j\alpha_j,$$

para cada $j = 0, 1, \ldots, n$. Esto se resuelve fácilmente dado que

$$a_j = \frac{1}{\alpha_j}\int_a^b w(x)f(x)\phi_j(x)\,dx.$$

La palabra "ortogonal" significa "en ángulo recto". Por lo que, en un sentido, las funciones ortogonales son perpendiculares entre sí.

Por lo tanto, el problema de aproximación de mínimos cuadrados se simplifica en gran medida cuando se seleccionan las funciones $\phi_0, \phi_1, \ldots, \phi_n$ para satisfacer la condición de *ortogonalidad* en la ecuación (8.7). El resto de esta sección está dedicado a estudiar conjuntos de este tipo.

Definición 8.5 Se dice que $\{\phi_0, \phi_1, \ldots, \phi_n\}$ es un **conjunto ortogonal de funciones** en el intervalo $[a, b]$ respecto a la función de peso w si

$$\int_a^b w(x)\phi_k(x)\phi_j(x)\,dx = \begin{cases} 0, & \text{cuando } j \neq k, \\ \alpha_j > 0, & \text{cuando } j = k. \end{cases}$$

Si, además, $\alpha_j = 1$ para cada $j = 0, 1, \ldots, n$, se dice que el conjunto es **ortonormal**. ■

Esta definición, junto con las observaciones anteriores, produce el siguiente teorema.

Teorema 8.6 Si $\{\phi_0, \ldots, \phi_n\}$ es un conjunto ortogonal de funciones en un intervalo $[a, b]$ respecto a la función de peso w, entonces la aproximación por mínimos cuadrados para f en $[a, b]$ respecto a w es

$$P(x) = \sum_{j=0}^{n} a_j\phi_j(x),$$

donde, para cada $j = 0, 1, \ldots, n$,

$$a_j = \frac{\int_a^b w(x)\phi_j(x)f(x)\,dx}{\int_a^b w(x)[\phi_j(x)]^2\,dx} = \frac{1}{\alpha_j}\int_a^b w(x)\phi_j(x)f(x)\,dx.$$ ■

A pesar de que la definición 8.5 y el teorema 8.6 permiten clases amplias de funciones ortogonales, en esta sección sólo consideraremos conjuntos ortogonales de polinomios. El siguiente teorema, que está basado en el **proceso Gram-Schmidt**, describe cómo construir polinomios ortogonales en $[a, b]$ respecto a la función de peso w.

Teorema 8.7 El conjunto de funciones polinomiales $\{\phi_0, \phi_1, \ldots, \phi_n\}$ definido de la siguiente forma es ortogonal en $[a, b]$ respecto a la función de peso w:

$$\phi_0(x) \equiv 1 \qquad \phi_1(x) = x - B_1, \text{para cada } x \text{ en } [a, b],$$

donde

$$B_1 = \frac{\int_a^b x w(x)[\phi_0(x)]^2\, dx}{\int_a^b w(x)[\phi_0(x)]^2\, dx},$$

y cuando $k \geq 2$,

$$\phi_k(x) = (x - B_k)\phi_{k-1}(x) - C_k\phi_{k-2}(x), \text{ por cada } x \text{ en } [a, b]$$

donde

$$B_k = \frac{\int_a^b x w(x)[\phi_{k-1}(x)]^2\, dx}{\int_a^b w(x)[\phi_{k-1}(x)]^2\, dx}$$

y

$$C_k = \frac{\int_a^b x w(x)\phi_{k-1}(x)\phi_{k-2}(x)\, dx}{\int_a^b w(x)[\phi_{k-2}(x)]^2\, dx}.$$

■

Erhard Schmidt (1876–1959) recibió su doctorado bajo la supervisión de David Hilbert, en 1905, para un problema relacionado con ecuaciones integrales. Schmidt publicó en 1907 un artículo en el que proporcionaba lo que ahora se conoce como proceso Gram-Schmidt para construir una base ortonormal para un conjunto de funciones. Éste generalizaba los resultados de Jorgen Pedersen Gram (1850–1916), quien consideró este problema al estudiar los mínimos cuadrados. Sin embargo, Laplace presentó un proceso similar mucho antes que Gram y que Schmidt.

El teorema 8.7 proporciona un procedimiento recursivo para construir un conjunto de polinomios ortonormales. La prueba de este teorema se sigue al aplicar inducción matemática al grado del polinomio $\phi_n(x)$.

Corolario 8.8 Para cualquier $n > 0$, el conjunto de funciones polinomiales $\{\phi_0, \ldots, \phi_n\}$ dado en el teorema 8.7 es linealmente independiente en $[a, b]$ y

$$\int_a^b w(x)\phi_n(x)Q_k(x)\, dx = 0,$$

para cualquier polinomio $Q_k(x)$ de grado $k < n$.

Demostración Para cada $k = 0, 1, \ldots, n$, $\phi_k(x)$ es un polinomio de grado k. Por lo que, el teorema 8.2 implica que $\{\phi_0, \ldots, \phi_n\}$ es un conjunto linealmente independiente.

Sea $Q_k(x)$ un polinomio de grado $k < n$. Mediante el teorema 8.3, existen números $c_0, \ldots, c_k$ de tal forma que

$$Q_k(x) = \sum_{j=0}^{k} c_j\phi_j(x).$$

Puesto que ϕ_n es ortogonal para ϕ_j para cada $j = 0, 1, \ldots, k$, tenemos

$$\int_a^b w(x)Q_k(x)\phi_n(x)\, dx = \sum_{j=0}^{k} c_j \int_a^b w(x)\phi_j(x)\phi_n(x)\, dx = \sum_{j=0}^{k} c_j \cdot 0 = 0.$$

■

Ilustración El conjunto de **polinomios de Legendre**, $\{P_n(x)\}$, es ortogonal en $[-1, 1]$ respecto a la función de peso $w(x) \equiv 1$. La definición clásica de los polinomios de Legendre requiere que $P_n(1) = 1$ para cada n y se utiliza una relación recursiva para generar los polinomios cuando $n \geq 2$. Esta normalización no será necesaria en nuestro análisis y los polinomios de aproximación por mínimos cuadrados generados en cualquier caso son fundamentalmente los mismos.

Usando el proceso de Gram-Schmidth con $P_0(x) \equiv 1$ obtenemos

$$B_1 = \frac{\int_{-1}^{1} x\, dx}{\int_{-1}^{1} dx} = 0 \quad \text{y} \quad P_1(x) = (x - B_1)P_0(x) = x.$$

Además,

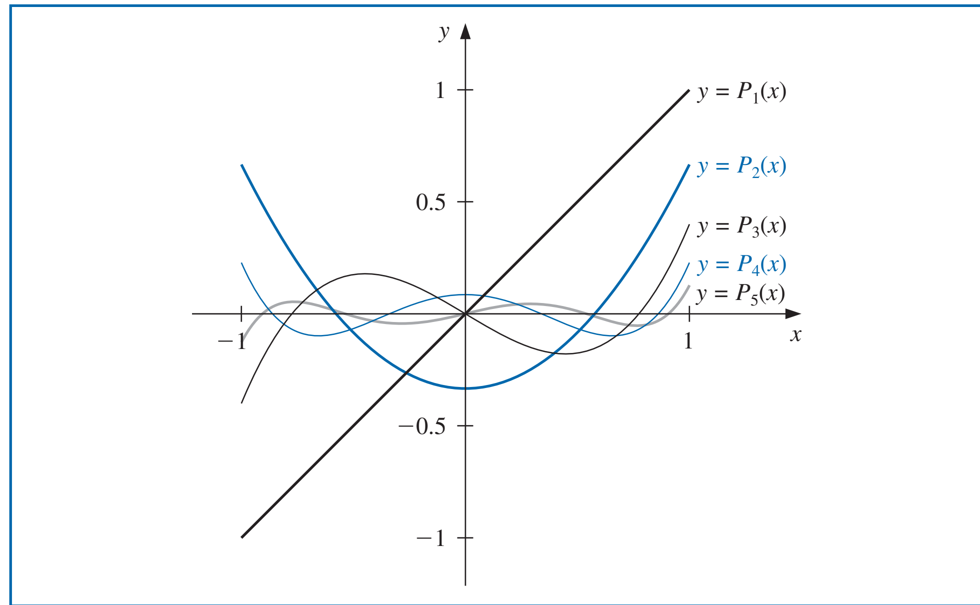
$$B_2 = \frac{\int_{-1}^{1} x^3\, dx}{\int_{-1}^{1} x^2\, dx} = 0 \quad \text{y} \quad C_2 = \frac{\int_{-1}^{1} x^2\, dx}{\int_{-1}^{1} 1\, dx} = \frac{1}{3},$$

por lo que

$$P_2(x) = (x - B_2)P_1(x) - C_2P_0(x) = (x-0)x - \frac{1}{3}\cdot 1 = x^2 - \frac{1}{3}.$$

Los polinomios de Legendre de grado superior que se muestran en la figura 8.9 se derivan de la misma forma. A pesar de que la integración puede ser tediosa, no es difícil con un sistema de álgebra para computadora.

Figura 8.9



Tenemos

$$P_3(x) = xP_2(x) - \frac{4}{15}P_1(x) = x^3 - \frac{1}{3}x - \frac{4}{15}x = x^3 - \frac{3}{5}x,$$

y los siguientes dos polinomios de Legendre son

$$P_4(x) = x^4 - \frac{6}{7}x^2 + \frac{3}{35} \quad \text{y} \quad P_5(x) = x^5 - \frac{10}{9}x^3 + \frac{5}{21}x.$$ ■

Los polinomios de Legendre fueron introducidos en la sección 4.7, en donde sus raíces, determinadas en la página 168, se utilizaron como los nodos en la cuadratura gaussiana.

La sección Conjunto de ejercicios 8.2 está disponible en línea. Encuentre la ruta de acceso en las páginas preliminares.

8.3 Polinomios de Chebyshev y ahorro de series de potencia

Pafnuty Lvovich Chebyshev (1821–1894) realizó un trabajo excepcional en muchas áreas como matemáticas aplicadas, teoría de números, teoría de aproximación y probabilidad. En 1852, viajó desde St. Petersburgo para visitar matemáticos en Francia, Inglaterra y Alemania. Lagrange y Legendre habían estudiado los conjuntos individuales de polinomios ortogonales, pero Chebyshev fue el primero en observar las consecuencias importantes de estudiar la teoría en general. Él desarrolló los polinomios de Chebyshev para estudiar la aproximación por mínimos cuadrados y la probabilidad y, después, aplicar sus resultados a la interpolación, cuadratura aproximada y otras áreas.

Los polinomios de Chebyshev $\{T_n(x)\}$ son ortogonales en $(-1, 1)$ respecto a la función de peso $w(x) = (1-x^2)^{-1/2}$. A pesar de que se pueden derivar con el método en la sección previa, es fácil proporcionar su definición y después mostrar que satisfacen las propiedades de ortogonalidad requeridas.

Para $x \in [-1, 1]$, defina

$$T_n(x) = \cos[n \arccos x], \qquad \text{para cada } n \geq 0. \tag{8.8}$$

Quizá a partir de esta definición no sea obvio que para cada n, $T_n(x)$ es un polinomio en x, pero ahora mostraremos esto. Primero, observe que

$$T_0(x) = \cos 0 = 1 \quad \text{y} \quad T_1(x) = \cos(\arccos x) = x.$$

Para $n \geq 1$, introducimos la sustitución $\theta = \arccos x$ para cambiar esta ecuación por

$$T_n(\theta(x)) \equiv T_n(\theta) = \cos(n\theta), \text{ donde } \theta \in [0, \pi].$$

Una relación de recurrencia se deriva al observar que

$$T_{n+1}(\theta) = \cos(n+1)\theta = \cos\theta\ \cos(n\theta) - \operatorname{sen}\theta\ \operatorname{sen}(n\theta)$$

y

$$T_{n-1}(\theta) = \cos(n-1)\theta = \cos\theta\ \cos(n\theta) + \operatorname{sen}\theta\ \operatorname{sen}(n\theta).$$

Al sumar estas ecuaciones obtenemos

$$T_{n+1}(\theta) = 2\cos\theta\ \cos(n\theta) - T_{n-1}(\theta).$$

Al regresar a la variable $x = \cos\theta$, tenemos, para $n \geq 1$,

$$T_{n+1}(x) = 2x\cos(n \arccos x) - T_{n-1}(x),$$

es decir,

$$T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x). \tag{8.9}$$

Puesto que $T_0(x) = 1$ y $T_1(x) = x$, la relación de recurrencia implica que los siguientes tres polinomios de Chebyshev son

$$T_2(x) = 2xT_1(x) - T_0(x) = 2x^2 - 1,$$

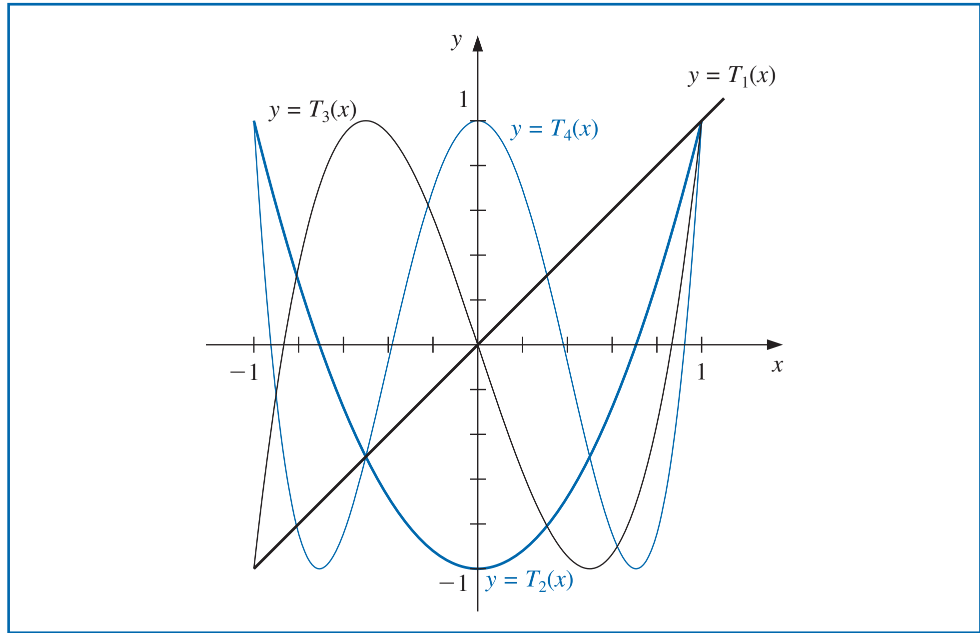
$$T_3(x) = 2xT_2(x) - T_1(x) = 4x^3 - 3x,$$

y

$$T_4(x) = 2xT_3(x) - T_2(x) = 8x^4 - 8x^2 + 1.$$

La relación de recurrencia también implica que cuando $n \geq 1$, $T_n(x)$ es un polinomio de grado n con coeficiente principal 2^{n-1}. Las gráficas de T_1, T_2, T_3 y T_4 se muestran en la figura 8.10.

Figura 8.10



Para mostrar la ortogonalidad de los polinomios de Chebyshev respecto a la función de peso $w(x) = (1 - x^2)^{-1/2}$, considere

$$\int_{-1}^{1} \frac{T_n(x)T_m(x)}{\sqrt{1-x^2}}\, dx = \int_{-1}^{1} \frac{\cos(n \arccos x)\cos(m \arccos x)}{\sqrt{1-x^2}}\, dx.$$

Al reintroducir la sustitución $\theta = \arccos x$ obtenemos

$$d\theta = -\frac{1}{\sqrt{1-x^2}}\, dx$$

y

$$\int_{-1}^{1} \frac{T_n(x)T_m(x)}{\sqrt{1-x^2}}\, dx = -\int_{\pi}^{0} \cos(n\theta)\cos(m\theta)\, d\theta = \int_{0}^{\pi} \cos(n\theta)\cos(m\theta)\, d\theta.$$

Suponga que $n \neq m$. Puesto que

$$\cos(n\theta)\cos(m\theta) = \frac{1}{2}[\cos(n+m)\theta + \cos(n-m)\theta],$$

tenemos

$$\int_{-1}^{1} \frac{T_n(x)T_m(x)}{\sqrt{1-x^2}}\, dx = \frac{1}{2}\int_{0}^{\pi} \cos((n+m)\theta)\, d\theta + \frac{1}{2}\int_{0}^{\pi} \cos((n-m)\theta)\, d\theta$$
$$= \left[\frac{1}{2(n+m)}\operatorname{sen}((n+m)\theta) + \frac{1}{2(n-m)}\operatorname{sen}((n-m)\theta)\right]_0^{\pi} = 0.$$

Mediante una técnica similar (consulte el ejercicio 11), también tenemos

$$\int_{-1}^{1} \frac{[T_n(x)]^2}{\sqrt{1-x^2}}\, dx = \frac{\pi}{2}, \quad \text{para cada } n \geq 1. \tag{8.10}$$

Los polinomios de Chebyshev se utilizan para minimizar el error de aproximación. Veremos cómo se usan para resolver dos problemas de este tipo:

- Una colocación óptima de puntos de interpolación para minimizar el error en la interpolación de Lagrange
- Un medio para reducir el grado de un polinomio de aproximación con pérdida mínima de precisión

El siguiente resultado afecta los ceros y los puntos extremos en $T_n(x)$.

Teorema 8.9 El polinomio de Chebyshev $T_n(x)$ de grado $n \geq 1$ tienen ceros simples en $[-1, 1]$ en

$$\bar{x}_k = \cos\left(\frac{2k-1}{2n}\pi\right), \quad \text{para cada } k = 1, 2, \ldots, n.$$

Además, $T_n(x)$ toma sus máximos absolutos en

$$\bar{x}'_k = \cos\left(\frac{k\pi}{n}\right) \quad \text{con} \quad T_n(\bar{x}'_k) = (-1)^k, \quad \text{para cada} \quad k = 0, 1, \ldots, n.$$

Demostración Si

$$\bar{x}_k = \cos\left(\frac{2k-1}{2n}\pi\right), \quad \text{para } k = 1, 2, \ldots, n.$$

Entonces

$$T_n(\bar{x}_k) = \cos(n \arccos \bar{x}_k) = \cos\left(n \arccos\left(\cos\left(\frac{2k-1}{2n}\pi\right)\right)\right) = \cos\left(\frac{2k-1}{2}\pi\right) = 0.$$

Pero $\bar{x}_k$ son distintas (consulte el ejercicio 12) y $T_n(x)$ es un polinomio de grado n, por lo que todos los ceros de $T_n(x)$ deben tener esta forma.

Para mostrar la segunda declaración, primero observe que

$$T'_n(x) = \frac{d}{dx}[\cos(n \arccos x)] = \frac{n \operatorname{sen}(n \arccos x)}{\sqrt{1-x^2}}$$

y que, cuando $k = 1, 2, \ldots, n-1$,

$$T'_n(\bar{x}'_k) = \frac{n \operatorname{sen}\left(n \arccos\left(\cos\left(\frac{k\pi}{n}\right)\right)\right)}{\sqrt{1-\left[\cos\left(\frac{k\pi}{n}\right)\right]^2}} = \frac{n \operatorname{sen}(k\pi)}{\operatorname{sen}\left(\frac{k\pi}{n}\right)} = 0.$$

Puesto que $T_n(x)$ es un polinomio de grado n, su derivada $T'_n(x)$ es un polinomio de grado $(n-1)$, y todos los ceros de $T'_n(x)$ se presentan en estos puntos distintos $n-1$ (que son diferentes se considera en el ejercicio 13). Las únicas otras posibilidades para los máximos de $T_n(x)$ se presentan en los extremos del intervalo $[-1, 1]$, es decir, en $\bar{x}'_0 = 1$ y en $\bar{x}'_n = -1$.

Para cualquier $k = 0, 1, \ldots, n$, tenemos

$$T_n(\bar{x}'_k) = \cos\left(n \arccos\left(\cos\left(\frac{k\pi}{n}\right)\right)\right) = \cos(k\pi) = (-1)^k.$$

Por lo que se presenta un máximo en cada valor par de k y un mínimo en cada valor impar. ■

Los polinomios mónicos de Chebyshev (polinomios con coeficiente principal 1) $\tilde{T}_n(x)$ se derivan a partir de los polinomios de Chebyshev $T_n(x)$ al dividir el coeficiente principal 2^{n-1}. Por lo tanto,

$$\tilde{T}_0(x) = 1 \quad \text{y} \quad \tilde{T}_n(x) = \frac{1}{2^{n-1}}T_n(x), \quad \text{por cada } n \geq 1. \tag{8.11}$$

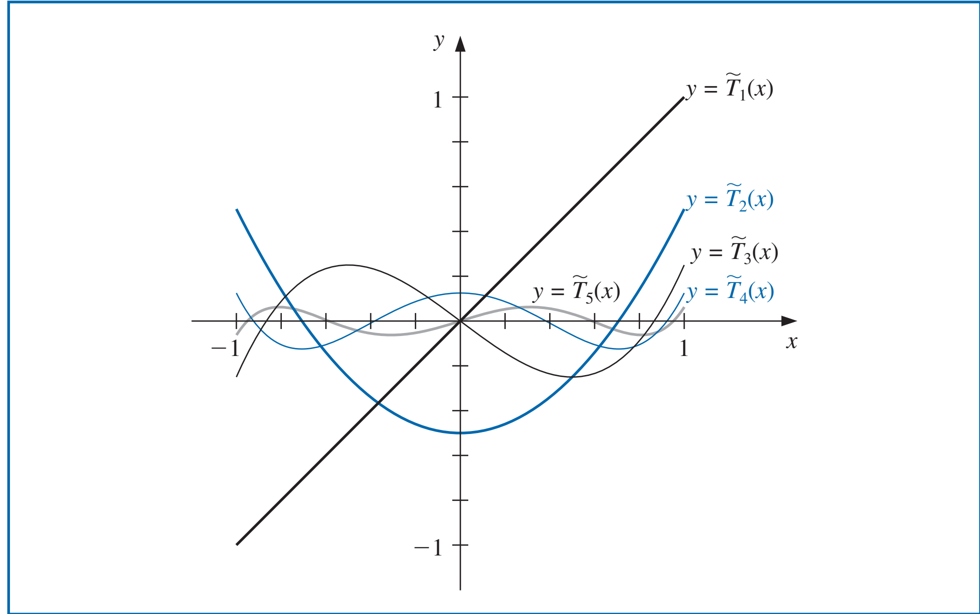
La relación de recurrencia satisfecha por los polinomios de Chebyshev implica que

$$\tilde{T}_2(x) = x\tilde{T}_1(x) - \frac{1}{2}\tilde{T}_0(x) \text{ y}$$

$$\tilde{T}_{n+1}(x) = x\tilde{T}_n(x) - \frac{1}{4}\tilde{T}_{n-1}(x), \text{ por cada } n \geq 2. \tag{8.12}$$

Las gráficas de $\tilde{T}_1$, $\tilde{T}_2$, $\tilde{T}_3$, $\tilde{T}_4$ y $\tilde{T}_5$ se muestran en la figura 8.11.

Figura 8.11



Puesto que $\tilde{T}_n(x)$ es sólo un múltiplo de $T_n(x)$, el teorema 8.9 implica que los ceros de $\tilde{T}_n(x)$ también se presentan en

$$\bar{x}_k = \cos\left(\frac{2k-1}{2n}\pi\right), \quad \text{para cada } k = 1, 2, \ldots, n,$$

y los valores extremos de $\tilde{T}_n(x)$ para $n \geq 1$, ocurre en

$$\bar{x}'_k = \cos\left(\frac{k\pi}{n}\right), \quad \text{con} \quad \tilde{T}_n(\bar{x}'_k) = \frac{(-1)^k}{2^{n-1}}, \quad \text{para cada } k = 0, 1, 2, \ldots, n. \tag{8.13}$$

Sea que $\tilde{\prod}_n$ denota **el conjunto de todos los polinomios mónicos de grado n**. La relación expresada en la ecuación (8.13) conduce a una propiedad de minimización importante que distingue $\tilde{T}_n(x)$ de los otros miembros de $\tilde{\prod}_n$.

Teorema 8.10 Los polinomios de la forma $\tilde{T}_n(x)$ cuando $n \geq 1$, tienen la propiedad de que

$$\frac{1}{2^{n-1}} = \max_{x\in[-1,1]} |\tilde{T}_n(x)| \leq \max_{x\in[-1,1]} |P_n(x)|, \quad \text{para toda } P_n(x) \in \tilde{\prod}_n.$$

Además, la igualdad se presenta sólo si $P_n \equiv \tilde{T}_n$.

Demostración Suponga que $P_n(x) \in \widetilde{\prod}_n$ y que

$$\max_{x\in[-1,1]} |P_n(x)| \le \frac{1}{2^{n-1}} = \max_{x\in[-1,1]} |\tilde{T}_n(x)|.$$

Sea $Q = \tilde{T}_n - P_n$. Entonces tanto $\tilde{T}_n(x)$ como $P_n(x)$ son polinomios mónicos de grado n, por lo que $Q(x)$ es un polinomio de grado máximo $(n - 1)$. Además, en los $n + 1$ puntos extremos $\bar{x}'_k$ de $\tilde{T}_n(x)$, tenemos

$$Q(\bar{x}'_k) = \tilde{T}_n(\bar{x}'_k) - P_n(\bar{x}'_k) = \frac{(-1)^k}{2^{n-1}} - P_n(\bar{x}'_k).$$

Sin embargo,

$$|P_n(\bar{x}'_k)| \le \frac{1}{2^{n-1}}, \quad \text{para cada } k = 0, 1, \ldots, n,$$

por lo que tenemos

$$Q(\bar{x}'_k) \le 0, \quad \text{cuando } k \text{ es impar} \quad \text{y} \quad Q(\bar{x}'_k) \ge 0, \quad \text{cuando } k \text{ es par}.$$

Puesto que Q es continuo, el teorema de valor intermedio implica que para cada $j = 0, 1, \ldots, n - 1$, el polinomio $Q(x)$ tiene por lo menos un cero entre $\bar{x}'_j$ y $\bar{x}'_{j+1}$. Por lo tanto, Q tiene por lo menos n ceros en el intervalo $[-1, 1]$. Pero el grado de $Q(x)$ es menor que n, por lo que $Q \equiv 0$. Esto implica que $P_n \equiv \tilde{T}_n$.

■

Minimización del error en la interpolación de Lagrange

El teorema 8.10 se puede utilizar para responder la pregunta de cuándo colocar nodos interpolantes para minimizar el error en la interpolación de Lagrange. El teorema 3.3 en la página 83 aplicado al intervalo $[-1, 1]$ establece que, si $x_0, \ldots, x_n$ son números distintos en el intervalo $[-1, 1]$ y si $f \in C^{n+1}[-1, 1]$, entonces, para cada $x \in [-1, 1]$, existe un número $\xi(x)$ en $(-1, 1)$ con

$$f(x) - P(x) = \frac{f^{(n+1)}(\xi(x))}{(n+1)!}(x - x_0)(x - x_1)\cdots(x - x_n),$$

donde $P(x)$ es el polinomio de interpolación de Lagrange. En general, no existe control sobre $\xi(x)$, por lo que minimizar el error mediante la colocación acertada de los nodos $x_0, \ldots, x_n$, seleccionamos $x_0, \ldots, x_n$ para minimizar la cantidad

$$|(x - x_0)(x - x_1)\cdots(x - x_n)|$$

a lo largo del intervalo $[-1, 1]$.

Puesto que $(x - x_0)(x - x_1)\cdots(x - x_n)$ es un polinomio mónico de grado $(n + 1)$, acabamos de observar que el mínimo se obtiene cuando

$$(x - x_0)(x - x_1)\cdots(x - x_n) = \tilde{T}_{n+1}(x).$$

El valor máximo de $|(x - x_0)(x - x_1)\cdots(x - x_n)|$ es más pequeño cuando se selecciona x_k para cada $k = 0, 1, \ldots, n$ es el $(k + 1)$-ésimo cero de $\tilde{T}_{n+1}$. Por lo tanto seleccionamos x_k como

$$\bar{x}_{k+1} = \cos\left(\frac{2k+1}{2(n+1)}\pi\right).$$

Puesto que $\max_{x\in[-1,1]} |\tilde{T}_{n+1}(x)| = 2^{-n}$, esto también implica que

$$\frac{1}{2^n} = \max_{x\in[-1,1]} |(x - \bar{x}_1)\cdots(x - \bar{x}_{n+1})| \le \max_{x\in[-1,1]} |(x - x_0)\cdots(x - x_n)|,$$

para cualquier selección de $x_0, x_1, \ldots, x_n$ en el intervalo $[-1, 1]$. El siguiente corolario sigue estas observaciones.

Corolario 8.11 Suponga que $P(x)$ es el polinomio de interpolación de grado a lo sumo n con nodos en los ceros de $T_{n+1}(x)$. Entonces

$$\max_{x\in[-1,1]} |f(x) - P(x)| \leq \frac{1}{2^n(n+1)!} \max_{x\in[-1,1]} |f^{(n+1)}(x)|, \quad \text{para cada } f \in C^{n+1}[-1,1]. \blacksquare$$

Minimización del error de aproximación en intervalos arbitrarios

La técnica para seleccionar puntos para minimizar el error de interpolación se amplía hasta un intervalo cerrado general $[a, b]$ al utilizar el cambio de variables

$$\tilde{x} = \frac{1}{2}[(b-a)x + a + b]$$

para transformar los números $\bar{x}_k$ en el intervalo $[-1, 1]$ en el número correspondiente $\bar{x}_k$ en el intervalo $[a, b]$, como se muestra en el siguiente ejemplo.

Ejemplo 1 Sea $f(x) = xe^x$ en $[0, 1.5]$. Compare los valores determinados por el polinomio de Lagrange con cuatro nodos igualmente espaciados con los dados por el polinomio de Lagrange con nodos determinados por los ceros del cuarto polinomio de Chebyshev.

Solución Los nodos igualmente espaciados $x_0 = 0$, $x_1 = 0.5$, $x_2 = 1$, y $x_3 = 1.5$ dan

$$\begin{aligned} L_0(x) &= -1.3333x^3 + 4.0000x^2 - 3.6667x + 1, \\ L_1(x) &= 4.0000x^3 - 10.000x^2 + 6.0000x, \\ L_2(x) &= -4.0000x^3 + 8.0000x^2 - 3.0000x, \quad \text{y} \\ L_3(x) &= 1.3333x^3 - 2.000x^2 + 0.66667x, \end{aligned}$$

que produce el polinomio

$$\begin{aligned} P_3(x) &= L_0(x)(0) + L_1(x)(0.5e^{0.5}) + L_2(x)e^1 + L_3(x)(1.5e^{1.5}) \\ &= 1.3875x^3 + 0.057570x^2 + 1.2730x. \end{aligned}$$

Para el segundo polinomio de interpolación, cambiamos los ceros $\bar{x}_k = \cos((2k+1)/8)\pi$, para $k = 0, 1, 2, 3$, de $\tilde{T}_4$, desde $[-1, 1]$ hasta $[0, 1.5]$,por medio de la transformación lineal

$$\tilde{x}_k = \frac{1}{2}[(1.5-0)\bar{x}_k + (1.5+0)] = 0.75 + 0.75\bar{x}_k.$$

Puesto que

$$\bar{x}_0 = \cos\frac{\pi}{8} = 0.92388, \ \bar{x}_1 = \cos\frac{3\pi}{8} = 0.38268, \ \bar{x}_2 = \cos\frac{5\pi}{8} = -0.38268, \ \text{y}$$
$$\bar{x}_3 = \cos\frac{7\pi}{8} = -0.92388,$$

tenemos

$$\tilde{x}_0 = 1.44291, \quad \tilde{x}_1 = 1.03701, \quad \tilde{x}_2 = 0.46299, \quad \text{y} \quad \tilde{x}_3 = 0.05709.$$

Los coeficientes polinomiales de Lagrange para este conjunto de nodos son

$$\begin{aligned} \tilde{L}_0(x) &= 1.8142x^3 - 2.8249x^2 + 1.0264x - 0.049728, \\ \tilde{L}_1(x) &= -4.3799x^3 + 8.5977x^2 - 3.4026x + 0.16705, \\ \tilde{L}_2(x) &= 4.3799x^3 - 11.112x^2 + 7.1738x - 0.37415, \quad \text{y} \\ \tilde{L}_3(x) &= -1.8142x^3 + 5.3390x^2 - 4.7976x + 1.2568. \end{aligned}$$

Los valores funcionales requeridos para estos polinomios se dan en las últimas dos columnas de la tabla 8.7. El polinomio de interpolación de grado máximo 3 es

$$\tilde{P}_3(x) = 1.3811x^3 + 0.044652x^2 + 1.3031x - 0.014352.$$

Tabla 8.7

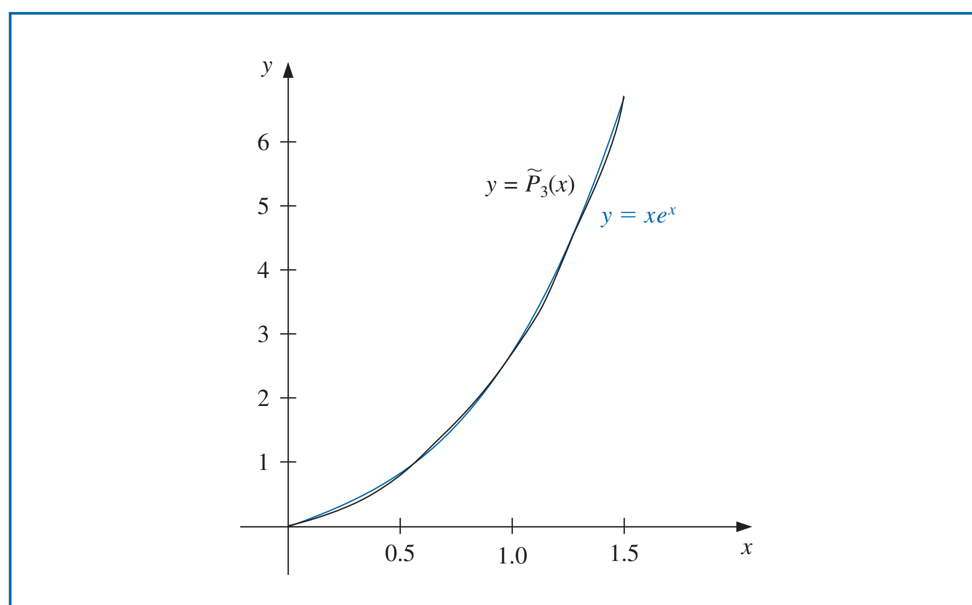
x	$f(x) = xe^x$	$\tilde{x}$	$f(\tilde{x}) = xe^x$
$x_0 = 0.0$	0.00000	$\tilde{x}_0 = 1.44291$	6.10783
$x_1 = 0.5$	0.824361	$\tilde{x}_1 = 1.03701$	2.92517
$x_2 = 1.0$	2.71828	$\tilde{x}_2 = 0.46299$	0.73560
$x_3 = 1.5$	6.72253	$\tilde{x}_3 = 0.05709$	0.060444

Para comparación, en la tabla 8.8 se listan varios valores de x, junto con los valores de $f(x)$, $P_3(x)$ y $\tilde{P}_3(x)$. A partir de esta tabla se puede observar que, a pesar de que el error por medio de $P_3(x)$ es menor al que resulta de utilizar $\tilde{P}_3(x)$ alrededor de la mitad de la tabla, el error máximo relacionado con el uso de $\tilde{P}_3(x)$, 0.0180, es considerablemente menor cuando se utiliza $P_3(x)$, lo cual da el error de 0.0290 (véase la figura 8.2). ■

Tabla 8.8

x	$f(x) = xe^x$	$P_3(x)$	$\lvert xe^x - P_3(x)\rvert$	$\tilde{P}_3(x)$	$\lvert xe^x - \tilde{P}_3(x)\rvert$
0.15	0.1743	0.1969	0.0226	0.1868	0.0125
0.25	0.3210	0.3435	0.0225	0.3358	0.0148
0.35	0.4967	0.5121	0.0154	0.5064	0.0097
0.65	1.245	1.233	0.012	1.231	0.014
0.75	1.588	1.572	0.016	1.571	0.017
0.85	1.989	1.976	0.013	1.974	0.015
1.15	3.632	3.650	0.018	3.644	0.012
1.25	4.363	4.391	0.028	4.382	0.019
1.35	5.208	5.237	0.029	5.224	0.016

Figura 8.12



Reducción del grado de los polinomios de aproximación

Los polinomios de Chebyshev también se pueden utilizar para reducir el grado de un polinomio de aproximación con una pérdida mínima de precisión. Puesto que los polinomios de Chebyshev tienen un valor absoluto máximo-mínimo que se distribuye de manera uniforme en un intervalo, se pueden utilizar para reducir el grado de un polinomio de aproximación sin exceder la tolerancia del error.

Considere aproximar un polinomio arbitrario de enésimo grado.

$$P_n(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0$$

en $[-1, 1]$ con un polinomio de grado como máximo $n - 1$. El objetivo es seleccionar $P_{n-1}(x)$ en $\prod_{n-1}$, de tal forma que

$$\max_{x\in[-1,1]} |P_n(x) - P_{n-1}(x)|$$

sea tan pequeño como resulte posible.

Primero observamos que $(P_n(x) - P_{n-1}(x))/a_n$ es un polinomio mónico de grado n, por lo que al aplicar el teorema 8.10 obtenemos

$$\max_{x\in[-1,1]} \left| \frac{1}{a_n}(P_n(x) - P_{n-1}(x)) \right| \geq \frac{1}{2^{n-1}}.$$

La igualdad se presenta precisamente cuando

$$\frac{1}{a_n}(P_n(x) - P_{n-1}(x)) = \tilde{T}_n(x).$$

Esto significa que deberíamos seleccionar

$$P_{n-1}(x) = P_n(x) - a_n \tilde{T}_n(x),$$

y con esta selección tenemos el valor mínimo

$$\max_{x\in[-1,1]} |P_n(x) - P_{n-1}(x)| = |a_n| \max_{x\in[-1,1]} \left| \frac{1}{a_n}(P_n(x) - P_{n-1}(x)) \right| = \frac{|a_n|}{2^{n-1}}.$$

Ilustración La función $f(x) = e^x$ se aproxima en el intervalo $[-1, 1]$ mediante el cuarto polinomio de Maclaurin

$$P_4(x) = 1 + x + \frac{x^2}{2} + \frac{x^3}{6} + \frac{x^4}{24},$$

que tiene error de truncamiento

$$|R_4(x)| = \frac{|f^{(5)}(\xi(x))||x^5|}{120} \leq \frac{e}{120} \approx 0.023, \quad \text{para } -1 \leq x \leq 1.$$

Suponga que un error de 0.05 es tolerable y que nos gustaría reducir el grado del polinomio de aproximación mientras nos mantenemos dentro de esta cota.

El polinomio de grado 3 o menor, que mejor se aproxima de manera uniforme a $P_4(x)$ en $[-1, 1]$

$$\begin{aligned} P_3(x) = P_4(x) - a_4\tilde{T}_4(x) &= 1 + x + \frac{x^2}{2} + \frac{x^3}{6} + \frac{x^4}{24} - \frac{1}{24}\left(x^4 - x^2 + \frac{1}{8}\right) \\ &= \frac{191}{192} + x + \frac{13}{24}x^2 + \frac{1}{6}x^3. \end{aligned}$$

Con esta selección, tenemos

$$|P_4(x) - P_3(x)| = |a_4\tilde{T}_4(x)| \leq \frac{1}{24} \cdot \frac{1}{2^3} = \frac{1}{192} \leq 0.0053.$$

Al sumar esta cota de error a la cota del error de truncamiento de Maclaurin obtenemos

$$0.023 + 0.0053 = 0.0283,$$

que está dentro del error permisible de 0.05.

El polinomio de grado 2 o menor, que mejor se aproxima de manera uniforme a P_3 (x) en $[-1, 1]$

$$P_2(x) = P_3(x) - \frac{1}{6}\tilde{T}_3(x)$$

$$= \frac{191}{192} + x + \frac{13}{24}x^2 + \frac{1}{6}x^3 - \frac{1}{6}\left(x^3 - \frac{3}{4}x\right) = \frac{191}{192} + \frac{9}{8}x + \frac{13}{24}x^2.$$

Sin embargo,

$$|P_3(x) - P_2(x)| = \left|\frac{1}{6}\tilde{T}_3(x)\right| = \frac{1}{6}\left(\frac{1}{2}\right)^2 = \frac{1}{24} \approx 0.042,$$

que, cuando se suma a la cota del error ya acumulada de 0.0283, excede la tolerancia de 0.05. Por consiguiente, el polinomio de menor grado que mejor se aproxima a e^x en $[-1, 1]$ con una cota de error menor a 0.05 es

$$P_3(x) = \frac{191}{192} + x + \frac{13}{24}x^2 + \frac{1}{6}x^3.$$

La tabla 8.9 muestra la función y los polinomios de aproximación en diferentes puntos en $[-1, 1]$. Observe que las entradas tabuladas para P_2 se encuentran dentro de la tolerancia de 0.05, aunque la cota de error para $P_2(x)$ excedió la tolerancia. ■

Tabla 8.9

x	e^x	$P_4(x)$	$P_3(x)$	$P_2(x)$	$\lvert e^x - P_2(x)\rvert$
–0.75	0.47237	0.47412	0.47917	0.45573	0.01664
–0.25	0.77880	0.77881	0.77604	0.74740	0.03140
0.00	1.00000	1.00000	0.99479	0.99479	0.00521
0.25	1.28403	1.28402	1.28125	1.30990	0.02587
0.75	2.11700	2.11475	2.11979	2.14323	0.02623

La sección Conjunto de ejercicios 8.3 está disponible en línea. Encuentre la ruta de acceso en las páginas preliminares.

8.4 Aproximación de función racional

La clase de polinomios algebraicos tiene algunas ventajas diferentes para uso en aproximación:

- Existe un número suficiente de polinomios para aproximar cualquier función continua en un intervalo cerrado dentro de una tolerancia arbitraria.

- Los polinomios se evalúan fácilmente en valores arbitrarios.
- Las derivadas e integrales de los polinomios existen y se determinan de manera sencilla.

La desventaja de utilizar polinomios para aproximación es su tendencia a la oscilación. Con frecuencia, esto causa cotas de error en la aproximación polinomial que exceden significativamente el error promedio de aproximación porque las cotas del error se determinan mediante el error máximo de aproximación. Ahora consideramos los métodos que distribuyen el error de aproximación de modo uniforme sobre el intervalo de aproximación. Estas técnicas implican funciones racionales.

Una **función racional** r de grado N tiene la forma

$$r(x) = \frac{p(x)}{q(x)},$$

donde $p(x)$ y $q(x)$ son polinomios cuyos grados suman N.

Todos los polinomios son una función racional (simplemente haga $q(x) \equiv 1$), por lo que la aproximación mediante funciones racionales proporciona resultados que no son peores que la aproximación por medio de polinomios. Sin embargo, las funciones racionales cuyo numerador y denominador tienen el mismo o casi el mismo grado, a menudo producen resultados de aproximación superiores a los métodos polinomiales para la misma cantidad de esfuerzo computacional. (Esta declaración está basada en la suposición de que la cantidad de esfuerzo computacional requerido para división es aproximadamente igual al de la multiplicación.)

Las funciones racionales tienen la ventaja añadida de permitir aproximación eficiente de funciones con discontinuidades infinitas cerca, pero fuera, del intervalo de aproximación. En general, la aproximación polinomial es inaceptable en esta situación.

Aproximación de Padé

Henri Padé (1863–1953) aportó un estudio sistemático de lo que hoy llamamos aproximaciones de Padé en su tesis doctoral en 1892. Probó los resultados en su estructura general y también estableció claramente la conexión entre las aproximaciones de Padé y fracciones continuas. Sin embargo, Daniel Bernoulli (1700–1782) y otros habían estudiado estas ideas desde 1730. James Stirling (1692–1770) presentó un método similar en *Methodus differentialis (Método diferencial)* publicado en el mismo año y Euler usó la aproximación de Padé para encontrar la suma de una serie.

Suponga que r es una función racional de grado $N = n + m$ de la forma

$$r(x) = \frac{p(x)}{q(x)} = \frac{p_0 + p_1 x + \cdots + p_n x^n}{q_0 + q_1 x + \cdots + q_m x^m},$$

que se utiliza para aproximar una función f en un intervalo cerrado I que contiene a cero. Para que r esté definida en cero, se requiere que $q_0 \neq 0$. De hecho, podemos suponer que $q_0 = 1$, si éste no es el caso, simplemente reemplazamos $p(x)$ por $p(x)/q_0$ y $q(x)$ por $q(x)/q_0$. Por consiguiente, existen $N + 1$ parámetros $q_1, q_2, \ldots, q_m, p_0, p_1, \ldots, p_n$ disponibles para la aproximación de f mediante r.

La **técnica de aproximación de Padé** es la extensión de la aproximación polinomial de Taylor para las funciones racionales. Se seleccionan los parámetros $N + 1$ de tal forma que $f^{(k)}(0) = r^{(k)}(0)$, para cada $k = 0, 1, \ldots, N$. Cuando $n = N$ y $m = 0$, la aproximación de Padé es simplemente el enésimo polinomio de Maclaurin.

Considere la diferencia

$$f(x) - r(x) = f(x) - \frac{p(x)}{q(x)} = \frac{f(x)q(x) - p(x)}{q(x)} = \frac{f(x)\sum_{i=0}^{m} q_i x^i - \sum_{i=0}^{n} p_i x^i}{q(x)}$$

y suponga que f tiene la expansión de la serie de Maclaurin $f(x) = \sum_{i=0}^{\infty} a_i x^i$. Entonces

$$f(x) - r(x) = \frac{\sum_{i=0}^{\infty} a_i x^i \sum_{i=0}^{m} q_i x^i - \sum_{i=0}^{n} p_i x^i}{q(x)}. \tag{8.14}$$

El objetivo es seleccionar las constantes $q_1, q_2, \ldots, q_m$ y $p_0, p_1, \ldots, p_n$ de tal forma que

$$f^{(k)}(0) - r^{(k)}(0) = 0, \text{ para cada } k = 0, 1, \ldots, N.$$

En la sección 2.4 (consulte, en particular, el ejercicio 10 en el Conjunto de ejercicios 2.4 en línea) encontramos que esto es equivalente a $f - r$ que tiene un cero de multiplicidad $N + 1$ en $x = 0$. Como consecuencia, seleccionamos $q_1, q_2, \ldots, q_m$ y $p_0, p_1, \ldots, p_n$ por lo que el numerador en el lado derecho de la ecuación (8.14),

$$(a_0 + a_1 x + \cdots)(1 + q_1 x + \cdots + q_m x^m) - (p_0 + p_1 x + \cdots + p_n x^n), \tag{8.15}$$

no tiene términos de grado menor o igual a N.

Para simplificar la notación, definimos $p_{n+1} = p_{n+2} = \cdots = p_N = 0$ y $q_{m+1} = q_{m+2} = \cdots = q_N = 0$. Ahora podemos expresar el coeficiente de x^k en la expresión (8.15) de manera más compacta como

$$\left(\sum_{i=0}^{k} a_i q_{k-i} \right) - p_k.$$

La función racional para la aproximación de Padé resulta a partir de la solución de $N + 1$ ecuaciones lineales

$$\sum_{i=0}^{k} a_i q_{k-i} = p_k, \quad k = 0, 1, \ldots, N$$

en las $N + 1$ incógnitas $q_1, q_2, \ldots, q_m, p_0, p_1, \ldots, p_n$.

Ejemplo 1 La expansión de la serie de Maclaurin para e^{-x} es

$$\sum_{i=0}^{\infty} \frac{(-1)^i}{i!} x^i.$$

Encuentre la aproximación de Padé para e^{-x} de grado 5 con $n = 3$ y $m = 2$.

Solución Para encontrar la aproximación de Padé, necesitamos seleccionar p_0, p_1, p_2, p_3, q_1, y q_2 por lo que los coeficientes de x^k para $k = 0, 1, \ldots, 5$ son 0 en la expresión

$$\left(1 - x + \frac{x^2}{2} - \frac{x^3}{6} + \cdots \right)(1 + q_1 x + q_2 x^2) - (p_0 + p_1 x + p_2 x^2 + p_3 x^3).$$

La expansión y agrupación de los términos produce

$$\begin{array}{llll}
x^5: & -\frac{1}{120} + \frac{1}{24} q_1 - \frac{1}{6} q_2 = 0; & x^2: & \frac{1}{2} - q_1 + q_2 = p_2; \\
x^4: & \frac{1}{24} - \frac{1}{6} q_1 + \frac{1}{2} q_2 = 0; & x^1: & -1 + q_1 = p_1; \\
x^3: & -\frac{1}{6} + \frac{1}{2} q_1 - q_2 = p_3; & x^0: & 1 = p_0.
\end{array}$$

La solución de este sistema es

$$\left\{ p_1 = -\frac{3}{5},\ p_2 = \frac{3}{20},\ p_3 = -\frac{1}{60},\ q_1 = \frac{2}{5},\ q_2 = \frac{1}{20} \right\}.$$

Por lo que la aproximación de Padé es

$$r(x) = \frac{1 - \frac{3}{5}x + \frac{3}{20}x^2 - \frac{1}{60}x^3}{1 + \frac{2}{5}x + \frac{1}{20}x^2}.$$

La tabla 8.10 muestra los valores de $r(x)$ y $P_5(x)$, el quinto polinomio de Maclaurin. La aproximación de Padé es claramente superior en este ejemplo. ■

Tabla 8.10

x	e^{-x}	$P_5(x)$	$\lvert e^{-x} - P_5(x)\rvert$	$r(x)$	$\lvert e^{-x} - r(x)\rvert$
0.2	0.81873075	0.81873067	8.64×10^{-8}	0.81873075	7.55×10^{-9}
0.4	0.67032005	0.67031467	5.38×10^{-6}	0.67031963	4.11×10^{-7}
0.6	0.54881164	0.54875200	5.96×10^{-5}	0.54880763	4.00×10^{-6}
0.8	0.44932896	0.44900267	3.26×10^{-4}	0.44930966	1.93×10^{-5}
1.0	0.36787944	0.36666667	1.21×10^{-3}	0.36781609	6.33×10^{-5}

El algoritmo 8.1 implementa la técnica de aproximación de Padé.

ALGORITMO **8.1**

Aproximación Racional de Padé

Para obtener la aproximación

$$r(x) = \frac{p(x)}{q(x)} = \frac{\sum_{i=0}^{n} p_i x^i}{\sum_{j=0}^{m} q_j x^j}$$

para una función determinada $f(x)$:

ENTRADA enteros no negativos m y n.

SALIDA coeficientes $q_0, q_1, \ldots, q_m$ y $p_0, p_1, \ldots, p_n$.

Paso 1 Determine $N = m + n$.

Paso 2 Para $i = 0, 1, \ldots, N$ determine $a_i = \dfrac{f^{(i)}(0)}{i!}$.
(*Los coeficientes del polinomio de Maclaurin son $a_0, \ldots, a_N$, que sería la entrada en lugar de calcularlos.*)

Paso 3 Determine $q_0 = 1$;
$p_0 = a_0$.

Paso 4 Para $i = 1, 2, \ldots, N$ haga los pasos 5–10. (*Establezca un sistema lineal con matriz B.*)

Paso 5 Para $j = 1, 2, \ldots, i-1$
si $j \le n$ entonces haga $b_{i,j} = 0$.

Paso 6 Si $i \le n$ entonces haga $b_{i,i} = 1$.

Paso 7 Para $j = i+1, i+2, \ldots, N$ determine $b_{i,j} = 0$.

Paso 8 Para $j = 1, 2, \ldots, i$
si $j \le m$ entonces haga $b_{i,n+j} = -a_{i-j}$.

Paso 9 Para $j = n+i+1, n+i+2, \ldots, N$ determine $b_{i,j} = 0$.

Paso 10 Determina $b_{i,N+1} = a_i$.

(*Los pasos 11–22 resuelven el sistema lineal mediante pivoteo parcial.*)

Paso 11 Para $i = n+1, n+2, \ldots, N-1$ haga los pasos 12–18.

Paso 12 Sea k el entero más pequeño con y $i \le k \le N$ y $|b_{k,i}|$
$= \text{máx}_{i \le j \le N} |b_{j,i}|$.
(*Encuentre el elemento pivote.*)

Paso 13 Si $b_{k,i} = 0$ entonces SALIDA ("El sistema es singular");
PARE.

Paso 14 Si $k \neq i$ entonces (*Intercambia fila i y fila k.*)
para $j = i, i+1, \ldots, N+1$ determine

$$b_{COPY} = b_{i,j};$$
$$b_{i,j} = b_{k,j};$$
$$b_{k,j} = b_{COPY}.$$

Paso 15 Para $j = i+1, i+2, \ldots, N$ haga los pasos 16–18. (*Realice la eliminación.*)

Paso 16 Determine $xm = \dfrac{b_{j,i}}{b_{i,i}}$.

Paso 17 Para $k = i+1, i+2, \ldots, N+1$
determine $b_{j,k} = b_{j,k} - xm \cdot b_{i,k}$.

Paso 18 Determine $b_{j,i} = 0$.

Paso 19 Si $b_{N,N} = 0$ entonces SALIDA ("El sistema es singular");
PARE.

Paso 20 Si $m > 0$ entonces determine $q_m = \dfrac{b_{N,N+1}}{b_{N,N}}$. (*Inicia la sustitución regresiva.*)

Paso 21 Para $i = N-1, N-2, \ldots, n+1$ determine $q_{i-n} = \dfrac{b_{i,N+1} - \sum_{j=i+1}^{N} b_{i,j} q_{j-n}}{b_{i,i}}$.

Paso 22 Para $i = n, n-1, \ldots, 1$ determine $p_i = b_{i,N+1} - \sum_{j=n+1}^{N} b_{i,j} q_{j-n}$.

Paso 23 SALIDA $(q_0, q_1, \ldots, q_m, p_0, p_1, \ldots, p_n)$;
PARE. (*El procedimiento fue exitoso.*) ■

Aproximación de fracción continuada

Es interesante comparar el número de operaciones aritméticas requeridas para los cálculos de $P_5(x)$ y $r(x)$ en el ejemplo 1. Mediante multiplicación anidada, $P_5(x)$ se puede expresar como

$$P_5(x) = \left(\left(\left(\left(-\frac{1}{120}x + \frac{1}{24}\right)x - \frac{1}{6}\right)x + \frac{1}{2}\right)x - 1\right)x + 1.$$

Al suponer que los coeficientes de $1, x, x^2, x^3, x^4$ y x^5 se representan como decimales, un solo cálculo de $P_5(x)$ en forma anidada requiere cinco multiplicaciones y cinco sumas/restas.

Usando multiplicación anidada, $r(x)$ se expresa como

$$r(x) = \frac{\left(\left(-\frac{1}{60}x + \frac{3}{20}\right)x - \frac{3}{5}\right)x + 1}{\left(\frac{1}{20}x + \frac{2}{5}\right)x + 1},$$

por lo que un solo cálculo de $r(x)$ requiere cinco multiplicaciones, cinco sumas/restas y una división. Por lo tanto, el esfuerzo computacional parece favorecer la aproximación polinomial.

Sin embargo, al reexpresar $r(x)$ mediante división continua, podemos escribir

$$\begin{aligned} r(x) &= \frac{1 - \frac{3}{5}x + \frac{3}{20}x^2 - \frac{1}{60}x^3}{1 + \frac{2}{5}x + \frac{1}{20}x^2} \\ &= \frac{-\frac{1}{3}x^3 + 3x^2 - 12x + 20}{x^2 + 8x + 20} \\ &= -\frac{1}{3}x + \frac{17}{3} + \frac{\left(-\frac{152}{3}x - \frac{280}{3}\right)}{x^2 + 8x + 20} \\ &= -\frac{1}{3}x + \frac{17}{3} + \frac{-\frac{152}{3}}{\left(\frac{x^2+8x+20}{x+(35/19)}\right)} \end{aligned}$$

o

$$r(x) = -\frac{1}{3}x + \frac{17}{3} + \frac{-\frac{152}{3}}{\left(x + \frac{117}{19} + \frac{3125/361}{(x+(35/19))}\right)}. \tag{8.16}$$

Escrita de esta forma, un solo cálculo de $r(x)$ requiere una multiplicación, cinco sumas/restas y dos divisiones. Si la cantidad de cálculos requeridos para división es aproximadamente igual para la multiplicación, el esfuerzo computacional requerido para una evaluación del polinomio $P_5(x)$ excede significativamente el requerido para una evaluación de la función racional $r(x)$.

Usando fracciones continuadas para aproximación racional es un tema que se origina en los trabajos de Christopher Clavius (1537–1612). Por ejemplo, Euler, Lagrange y Hermite lo usaron en los siglos XVIII y XIX.

Al expresar una aproximación de función racional en una forma como la ecuación (8.16) recibe el nombre de **fracción continuada**. Ésta es una técnica de aproximación clásica de interés actual debido a la eficiencia computacional de su representación. Es, sin embargo, una técnica especializada que analizaremos más adelante. Un tratamiento bastante amplio de este tema y de la aproximación racional en general se puede encontrar en [RR], p. 285-322.

A pesar de que la aproximación de la función racional en el ejemplo 1 da resultados superiores a la aproximación polinomial del mismo grado observe que la aproximación tiene una amplia variación en precisión. La aproximación en 0.2 es precisa dentro de 8×10^{-9}, pero en 1.0 la aproximación y la función sólo concuerdan dentro de 7×10^{-5}. Se espera esta variación de precisión porque la aproximación de Padé está basada en una representación de e^{-x}, y la representación de Taylor tiene una amplia variación de precisión en [0.2, 1.0].

Aproximación de la función racional de Chebyshev

Para obtener aproximaciones de función racional precisa de forma más uniforme, usamos los polinomios de Chebyshev. El método general de aproximación de función racional de Chebyshev procede de la misma forma que la aproximación de Padé, excepto que cada término x^k en la aproximación de Padé se reemplaza por el polinomio de k-ésimo grado de Chebyshev $T_k(x)$.

Suponga que queremos aproximar la función f mediante una función racional r de enésimo grado escrita en la forma

$$r(x) = \frac{\sum_{k=0}^{n} p_k T_k(x)}{\sum_{k=0}^{m} q_k T_k(x)}, \quad \text{donde } N = n + m \text{ y } q_0 = 1.$$

Al escribir $f(x)$ en una serie relacionada con los polinomios de Chebyshev como

$$f(x) = \sum_{k=0}^{\infty} a_k T_k(x)$$

da

$$f(x) - r(x) = \sum_{k=0}^{\infty} a_k T_k(x) - \frac{\sum_{k=0}^{n} p_k T_k(x)}{\sum_{k=0}^{m} q_k T_k(x)}$$

o

$$f(x) - r(x) = \frac{\sum_{k=0}^{\infty} a_k T_k(x) \sum_{k=0}^{m} q_k T_k(x) - \sum_{k=0}^{n} p_k T_k(x)}{\sum_{k=0}^{m} q_k T_k(x)}. \tag{8.17}$$

Los coeficientes $q_1, q_2, \ldots, q_m$ y $p_0, p_1, \ldots, p_n$ se seleccionan de tal forma que el numerador en el lado derecho de esta ecuación tiene cero coeficientes para $T_k(x)$ cuando $k = 0, 1, \ldots, N$. Esto implica que la serie

$$(a_0T_0(x) + a_1T_1(x) + \cdots)(T_0(x) + q_1T_1(x) + \cdots + q_mT_m(x))$$
$$- (p_0T_0(x) + p_1T_1(x) + \cdots + p_nT_n(x))$$

no tiene términos de grado menor o igual a N.

Con el procedimiento de Chebyshev surgen dos problemas que lo hacen más difícil de implementar que el método de Padé. Uno se presenta porque el producto del polinomio $q(x)$ y la serie para $f(x)$ implican productos de los polinomios de Chebyshev. Este problema se resuelve al utilizar la relación

$$T_i(x)T_j(x) = \frac{1}{2}\left[T_{i+j}(x) + T_{|i-j|}(x)\right]. \tag{8.18}$$

(Consulte el ejercicio 10 de la sección 8.3.) El otro problema es más difícil de resolver e implica el cálculo de la serie de Chebyshev para $f(x)$. En teoría, esto no es difícil, si

$$f(x) = \sum_{k=0}^{\infty} a_k T_k(x),$$

entonces la ortogonalidad de los polinomios de Chebyshev implica que

$$a_0 = \frac{1}{\pi}\int_{-1}^{1} \frac{f(x)}{\sqrt{1-x^2}}\,dx \quad \text{y} \quad a_k = \frac{2}{\pi}\int_{-1}^{1} \frac{f(x)T_k(x)}{\sqrt{1-x^2}}\,dx, \quad \text{donde } k \geq 1.$$

Prácticamente, sin embargo, estas integrales rara vez se pueden evaluar de forma cerrada y se requiere una técnica de integración numérica para cada evaluación.

Ejemplo 2 Los primeros cinco términos de la expansión de Chebyshev para e^{-x} son

$$\tilde{P}_5(x) = 1.266066T_0(x) - 1.130318T_1(x) + 0.271495T_2(x) - 0.044337T_3(x)$$
$$+ 0.005474T_4(x) - 0.000543T_5(x).$$

Determine la aproximación racional de Chebyshev de grado 5 con $n = 3$ y $m = 2$.

Solución Encontrar esta aproximación requiere seleccionar p_0, p_1, p_2, p_3, q_1 y q_2 y de tal forma que $k = 0, 1, 2, 3, 4$ y 5, los coeficientes de $T_k(x)$ son 0 en la expansión

$$\tilde{P}_5(x)[T_0(x) + q_1T_1(x) + q_2T_2(x)] - [p_0T_0(x) + p_1T_1(x) + p_2T_2(x) + p_3T_3(x)].$$

Usando la relación (8.18) y agrupando términos se obtienen las ecuaciones

$$\begin{aligned}
T_0: &\quad 1.266066 - 0.565159q_1 + 0.1357485q_2 = p_0,\\
T_1: &\quad -1.130318 + 1.401814q_1 - 0.587328q_2 = p_1,\\
T_2: &\quad 0.271495 - 0.587328q_1 + 1.268803q_2 = p_2,\\
T_3: &\quad -0.044337 + 0.138485q_1 - 0.565431q_2 = p_3,\\
T_4: &\quad 0.005474 - 0.022440q_1 + 0.135748q_2 = 0, \quad \text{y}\\
T_5: &\quad -0.000543 + 0.002737q_1 - 0.022169q_2 = 0.
\end{aligned}$$

La solución de este sistema produce la función racional

$$r_T(x) = \frac{1.055265T_0(x) - 0.613016T_1(x) + 0.077478T_2(x) - 0.004506T_3(x)}{T_0(x) + 0.378331T_1(x) + 0.022216T_2(x)}.$$

Al inicio de la sección 8.3 encontramos que

$$T_0(x) = 1,\; T_1(x) = x,\; T_2(x) = 2x^2 - 1, \;\text{ y }\; T_3(x) = 4x^3 - 3x.$$

Al utilizarlas para convertir una expresión relacionada con potencias de x obtenemos

$$r_T(x) = \frac{0.977787 - 0.599499x + 0.154956x^2 - 0.018022x^3}{0.977784 + 0.378331x + 0.044432x^2}.$$

La tabla 8.11 contiene los valores de $r_T(x)$ y, para propósitos de comparación, los valores de $r(x)$ obtenidos en el ejemplo 1. Observe que la aproximación dada por $r(x)$ es superior al de $r_T(x)$ para $x = 0.02$ y 0.4 pero el error máximo para $r(x)$ es $6.33 = 10^{-5}$ en comparación con 9.13×10^{-6} para $r_T(x)$. ■

Tabla 8.11

x	e^{-x}	$r(x)$	$\lvert e^{-x} - r(x)\rvert$	$r_T(x)$	$\lvert e^{-x} - r_T(x)\rvert$
0.2	0.81873075	0.81873075	7.55×10^{-9}	0.81872510	5.66×10^{-6}
0.4	0.67032005	0.67031963	4.11×10^{-7}	0.67031310	6.95×10^{-6}
0.6	0.54881164	0.54880763	4.00×10^{-6}	0.54881292	1.28×10^{-6}
0.8	0.44932896	0.44930966	1.93×10^{-5}	0.44933809	9.13×10^{-6}
1.0	0.36787944	0.36781609	6.33×10^{-5}	0.36787155	7.89×10^{-6}

La aproximación de Chebyshev se puede generar mediante el algoritmo 8.2.

ALGORITMO **8.2**

Aproximación racional de Chebyshev

Para obtener la aproximación racional

$$r_T(x) = \frac{\sum_{k=0}^{n} p_k T_k(x)}{\sum_{k=0}^{m} q_k T_k(x)}$$

para una función determinada $f(x)$:

ENTRADA enteros no negativos m y n.

SALIDA los coeficientes $q_0, q_1, \ldots, q_m$ y $p_0, p_1, \ldots, p_n$.

Paso 1 Determine $N = m + n$.

Paso 2 Determine $a_0 = \dfrac{2}{\pi}\int_0^{\pi} f(\cos\theta)\,d\theta$; (*El coeficiente* a_0 *se duplica para eficiencia computacional*.)

Para $k = 1, 2, \ldots, N + m$ determine

$$a_k = \frac{2}{\pi}\int_0^{\pi} f(\cos\theta)\cos k\theta\,d\theta.$$

(*Las integrales se pueden evaluar usando el procedimiento de integración o los coeficientes se pueden colocar directamente*.)

Paso 3 Determine $q_0 = 1$.

Paso 4 Para $i = 0, 1, \ldots, N$ haga los pasos 5–9. (*Configura un sistema lineal con matriz B*.)

Paso 5 Para $j = 0, 1, \ldots, i$
si $j \leq n$ entonces determine $b_{i,j} = 0$.

Paso 6 Si $i \leq n$ entonces determine $b_{i,i} = 1$.

Paso 7 Para $j = i + 1, i + 2, \ldots, n$ determine $b_{i,j} = 0$.

Paso 8 Para $j = n + 1, n + 2, \ldots, N$
si $i \neq 0$ entonces determine $b_{i,j} = -\frac{1}{2}(a_{i+j-n} + a_{|i-j+n|})$
también determine $b_{i,j} = -\frac{1}{2}a_{j-n}$.

Paso 9 Si $i \neq 0$ entonces determine $b_{i,N+1} = a_i$
también determine $b_{i,N+1} = \frac{1}{2}a_i$.

(*Pasos 10–21 resuelven el sistema lineal mediante pivoteo parcial*.)

Paso 10 Para $i = n + 1, n + 2, \ldots, N - 1$ haga los pasos 11–17.

Paso 11 Sea k el entero más pequeño con $i \leq k \leq N$ y
$|b_{k,i}| = \text{máx}_{i \leq j \leq N} |b_{j,i}|$. (*Encuentre el elemento de pivote*.)

Paso 12 Si $b_{k,i} = 0$ entonces SALIDA ("El sistema es singular");
PARE.

Paso 13 Si $k \neq i$ entonces (*Intercambie la fila y y k*.)
para $j = i, i + 1, \ldots, N + 1$ determine

$$b_{COPY} = b_{i,j};$$
$$b_{i,j} = b_{k,j};$$
$$b_{k,j} = b_{COPY}.$$

Paso 14 Para $j = i + 1, i + 2, \ldots, N$ haga los pasos 15–17. (*Realice la eliminación*.)

Paso 15 Determine $xm = \dfrac{b_{j,i}}{b_{i,i}}$.

Paso 16 Para $k = i + 1, i + 2, \ldots, N + 1$
determine $b_{j,k} = b_{j,k} - xm \cdot b_{i,k}$.

Paso 17 Determine $b_{j,i} = 0$.

Paso 18 Si $b_{N,N} = 0$ entonces SALIDA ("En sistema es singular");
PARE.

Paso 19 Si $m > 0$ entonces determine $q_m = \frac{b_{N,N+1}}{b_{N,N}}$. (*Comience la sustitución regresiva.*)

Paso 20 Para $i = N-1, N-2, \ldots, n+1$ determine $q_{i-n} = \frac{b_{i,N+1} - \sum_{j=i+1}^{N} b_{i,j}q_{j-n}}{b_{i,i}}$.

Paso 21 Para $i = n, n-1, \ldots, 0$ determine $p_i = b_{i,N+1} - \sum_{j=n+1}^{N} b_{i,j}q_{j-n}$.

Paso 22 SALIDA $(q_0, q_1, \ldots, q_m, p_0, p_1, \ldots, p_n)$;
PARE. (*El procedimiento fue exitoso*.) ■

En 1930, Evgeny Remez (1896–1975) creó métodos computacionales generales de aproximación de Chebyshev para polinomios. Más adelante, desarrolló un algoritmo similar para la aproximación racional de funciones continuas definidas en un intervalo con un grado prescrito de precisión. Su trabajo abarcaba varias áreas de la teoría de aproximación, así como los métodos para aproximar las soluciones de ecuaciones diferenciales.

El método de Chebyshev no produce la mejor aproximación funcional racional en el sentido de la aproximación cuyo error máximo es mínimo. Sin embargo, es posible usar el método como punto de inicio para un método iterativo conocido como segundo algoritmo Remez, que converge en la mejor aproximación. Un análisis de las técnicas relacionadas con este procedimiento y una mejora de este algoritmo se puede encontrar en [RR], p. 292–305 o en [Pow], p. 90–92.

La sección Conjunto de ejercicios 8.4 está disponible en línea. Encuentre la ruta de acceso en las páginas preliminares.

8.5 Aproximación polinomial trigonométrica

El uso de series de funciones de seno y coseno para representar funciones arbitrarias tiene su inicios en la década de 1750 con el estudio del movimiento de una cuerda vibrante. Este problema fue considerado por Jean d'Alembert y, después, por el matemático más destacado de todos los tiempos, Leonhard Euler. Pero fue Daniel Bernoulli quien abogó primero por el uso de las sumas infinitas de seno y coseno como una solución para el problema, sumas que ahora conocemos como series de Fourier. En la primera parte del siglo XIX, Jean Baptiste Joseph Fourier utilizó estas series para estudiar el flujo de calor y desarrolló una teoría bastante compleja sobre el tema.

A finales del siglo XVII y principios del XVIII, la familia Bernoulli produjo no menos de ocho matemáticos y físicos destacados. El trabajo más importante de Daniel Bernoulli implicaba la presión, la densidad y la velocidad del flujo de fluido, que resultó en lo que se conoce como el *principio de Bernoulli.*

La primera observación en el desarrollo de la serie de Fourier es que, para cada entero positivo n, el conjunto de funciones $\{\phi_0, \phi_1, \ldots, \phi_{2n-1}\}$, donde

$$\phi_0(x) = \frac{1}{2},$$

$$\phi_k(x) = \cos kx, \quad \text{para cada } k = 1, 2, \ldots, n,$$

y

$$\phi_{n+k}(x) = \operatorname{sen} kx, \quad \text{para cada } k = 1, 2, \ldots, n-1,$$

es un conjunto ortogonal en $[-\pi, \pi]$ respecto a $w(x) \equiv 1$. Esta ortogonalidad sigue el hecho de que para cada entero j, las integrales de sen jx y cos jx sobre $[-\pi, \pi]$ son 0 y podemos reescribir los productos de las funciones seno y coseno como sumas al utilizar las tres identidades trigonométricas

$$\operatorname{sen} t_1 \operatorname{sen} t_2 = \frac{1}{2}[\cos(t_1 - t_2) - \cos(t_1 + t_2)],$$

$$\cos t_1 \cos t_2 = \frac{1}{2}[\cos(t_1 - t_2) + \cos(t_1 + t_2)], \quad \text{y} \tag{8.19}$$

$$\operatorname{sen} t_1 \cos t_2 = \frac{1}{2}[\operatorname{sen}(t_1 - t_2) + \operatorname{sen}(t_1 + t_2)].$$

Polinomios trigonométricos ortogonales

Sea $\mathcal{T}_n$ el conjunto de todas las combinaciones lineales de las funciones $\phi_0, \phi_1, \ldots, \phi_{2n-1}$. Este conjunto recibe el nombre de conjunto de **polinomios trigonométricos** de grado menor o igual a n. (Algunas fuentes también incluyen una función adicional en el conjunto $\phi_{2n}(x) = \operatorname{sen} nx$.)

Para una función $f \in C[-\pi, \pi]$, queremos encontrar la aproximación de mínimos cuadrados continuos mediante las funciones en $\mathcal{T}_n$ de la forma

$$S_n(x) = \frac{a_0}{2} + a_n \cos nx + \sum_{k=1}^{n-1}(a_k \cos kx + b_k \operatorname{sen} kx).$$

Puesto que el conjunto de funciones $\{\phi_0, \phi_1, \ldots, \phi_{2n-1}\}$ es ortogonal en $[-\pi, \pi]$ respecto a $w(x) \equiv 1$, se sigue del teorema 8.6 en la página 382 y las ecuaciones en (8.19) sobre que la selección adecuada de coeficientes es

$$a_k = \frac{\int_{-\pi}^{\pi} f(x) \cos kx\, dx}{\int_{-\pi}^{\pi} (\cos kx)^2\, dx} = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\cos kx\, dx, \quad \text{para cada } k = 0, 1, 2, \ldots, n, \tag{8.20}$$

y

$$b_k = \frac{\int_{-\pi}^{\pi} f(x) \operatorname{sen} kx\, dx}{\int_{-\pi}^{\pi} (\operatorname{sen} kx)^2\, dx} = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\operatorname{sen} kx\, dx, \quad \text{para cada } k = 1, 2, \ldots, n-1. \tag{8.21}$$

Joseph Fourier (1768–1830) publicó su teoría de series trigonométricas en *Théorie analytique de la chaleur (Teoría analítica del calor)* para resolver el problema de distribución de calor de estado estable en un sólido.

El límite de $S_n(x)$ cuando $n \to \infty$ recibe el nombre de **serie de Fourier** de f. Las series de Fourier se usan para describir la solución de las diferentes ecuaciones ordinarias y diferenciales parciales que se presentan en situaciones físicas.

Ejemplo 1 Determine el polinomio trigonométrico a partir de $\mathcal{T}_n$ que aproxima

$$f(x) = |x|, \quad \text{para } -\pi < x < \pi.$$

Solución Primero necesitamos los coeficientes

$$a_0 = \frac{1}{\pi}\int_{-\pi}^{\pi} |x|\, dx = -\frac{1}{\pi}\int_{-\pi}^{0} x\, dx + \frac{1}{\pi}\int_{0}^{\pi} x\, dx = \frac{2}{\pi}\int_{0}^{\pi} x\, dx = \pi,$$

$$a_k = \frac{1}{\pi}\int_{-\pi}^{\pi} |x| \cos kx\, dx = \frac{2}{\pi}\int_{0}^{\pi} x\cos kx\, dx = \frac{2}{\pi k^2}\left[(-1)^k - 1\right],$$

para cada $k = 1, 2, \ldots, n$, y

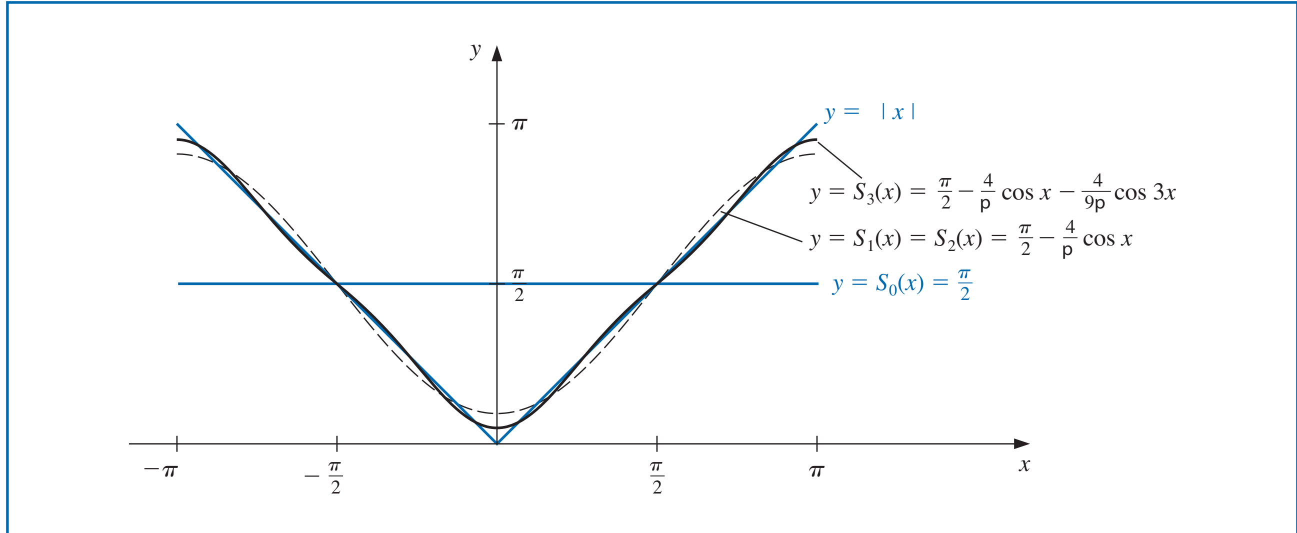
$$b_k = \frac{1}{\pi}\int_{-\pi}^{\pi} |x| \operatorname{sen} kx\, dx = 0, \quad \text{para cada } k = 1, 2, \ldots, n-1.$$

Al hecho de que todas las b_k's son 0 sigue que $g(x) = |x| \operatorname{sen} kx$ es una función impar para cada k y que la integral de una función impar continua sobre un intervalo de la forma $[-a, a]$ es 0 (consulte los ejercicios 15 y 16). El polinomio trigonométrico de $\mathcal{T}_n$ que se aproxima a f es, por lo tanto,

$$S_n(x) = \frac{\pi}{2} + \frac{2}{\pi}\sum_{k=1}^{n} \frac{(-1)^k - 1}{k^2}\cos kx.$$

Los primeros polinomios trigonométricos para $f(x) = |x|$ se muestran en la figura 8.13. ■

Figura 8.13



La serie de Fourier para f es

$$S(x) = \lim_{n\to\infty} S_n(x) = \frac{\pi}{2} + \frac{2}{\pi}\sum_{k=1}^{\infty}\frac{(-1)^k - 1}{k^2}\cos kx.$$

Puesto que $|\cos kx| \leq 1$ para cada k y x, la serie converge, y $S(x)$ existe para todos los números reales x.

Aproximación trigonométrica discreta

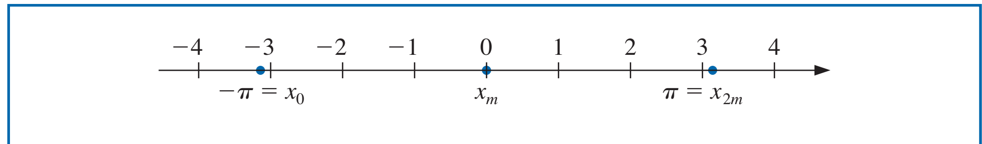
Existe un análogo discreto que es útil para la aproximación de *mínimos cuadrados discretos* y la interpolación de grandes cantidades de datos.

Suponga que se proporciona un conjunto de puntos de datos pares $2m$ $\{(x_j, y_j)\}_{j=0}^{2m-1}$, con los primeros elementos en pares que dividen de manera uniforme un intervalo cerrado. Por conveniencia, suponemos que el intervalo es $[-\pi, \pi]$; por lo que, como se muestra en la figura 8.14,

$$x_j = -\pi + \left(\frac{j}{m}\right)\pi, \quad \text{para cada } j = 0, 1, \ldots, 2m-1. \tag{8.22}$$

Si no es $[-\pi, \pi]$, se podría usar una transformación lineal simple para transformar los datos en esta forma.

Figura 8.14



La meta en el caso discreto es determinar el polinomio trigonométrico $S_n(x)$ en $\mathcal{T}_n$ que minimizará

$$E(S_n) = \sum_{j=0}^{2m-1}[y_j - S_n(x_j)]^2.$$

Para hacerlo, necesitamos seleccionar las constantes $a_0, a_1, \ldots, a_n, b_1, b_2, \ldots, b_{n-1}$ para minimizar

$$E(S_n) = \sum_{j=0}^{2m-1} \left\{ y_j - \left[\frac{a_0}{2} + a_n \cos nx_j + \sum_{k=1}^{n-1} (a_k \cos kx_j + b_k \operatorname{sen} kx_j) \right] \right\}^2. \quad (8.23)$$

La determinación de las constantes se simplifica mediante el hecho de que el conjunto $\{\phi_0, \phi_1, \ldots, \phi_{2n-1}\}$ es ortogonal respecto a la suma sobre los puntos uniformemente espaciados $\{x_j\}_{j=0}^{2m-1}$ en $[-\pi, \pi]$. Con esto queremos decir que para cada $k \neq l$,

$$\sum_{j=0}^{2m-1} \phi_k(x_j)\phi_l(x_j) = 0. \quad (8.24)$$

Para mostrar esta ortogonalidad utilizamos el siguiente lema.

Lema 8.12 Suponga que el entero r no es un múltiplo de $2m$. Entonces

- $$\sum_{j=0}^{2m-1} \cos rx_j = 0 \quad \text{y} \quad \sum_{j=0}^{2m-1} \operatorname{sen} rx_j = 0.$$

Además, si r no es un múltiplo de m, entonces

- $$\sum_{j=0}^{2m-1} (\cos rx_j)^2 = m \quad \text{y} \quad \sum_{j=0}^{2m-1} (\operatorname{sen} rx_j)^2 = m.$$

Euler usó primero el símbolo i en 1794 para representar $\sqrt{-1}$ en sus memorias *De formulis differentialibus Angularibus.*

Demostración La *fórmula de Euler* establece que con $i^2 = -1$, tenemos, para todos los números reales z,

$$e^{iz} = \cos z + i \operatorname{sen} z. \quad (8.25)$$

Al aplicar estos resultados obtenemos

$$\sum_{j=0}^{2m-1} \cos rx_j + i \sum_{j=0}^{2m-1} \operatorname{sen} rx_j = \sum_{j=0}^{2m-1} \big(\cos rx_j + i \operatorname{sen} rx_j\big) = \sum_{j=0}^{2m-1} e^{irx_j}.$$

Pero

$$e^{irx_j} = e^{ir(-\pi + j\pi/m)} = e^{-ir\pi} \cdot e^{irj\pi/m},$$

por lo que

$$\sum_{j=0}^{2m-1} \cos rx_j + i \sum_{j=0}^{2m-1} \operatorname{sen} rx_j = e^{-ir\pi} \sum_{j=0}^{2m-1} e^{irj\pi/m}.$$

Puesto que $\sum_{j=0}^{2m-1} e^{irj\pi/m}$ es una serie geométrica con el primer término 1 y radio $e^{ir\pi/m} \neq 1$, tenemos

$$\sum_{j=0}^{2m-1} e^{irj\pi/m} = \frac{1 - (e^{ir\pi/m})^{2m}}{1 - e^{ir\pi/m}} = \frac{1 - e^{2ir\pi}}{1 - e^{ir\pi/m}}.$$

Pero $e^{2ir\pi} = \cos 2r\pi + i \operatorname{sen} 2r\pi = 1$, por lo que $1 - e^{2ir\pi} = 0$ y

$$\sum_{j=0}^{2m-1} \cos rx_j + i \sum_{j=0}^{2m-1} \operatorname{sen} rx_j = e^{-ir\pi} \sum_{j=0}^{2m-1} e^{irj\pi/m} = 0.$$

Esto implica que tanto la parte real como la imaginaria son cero, por lo que

$$\sum_{j=0}^{2m-1} \cos r x_j = 0 \quad \text{y} \quad \sum_{j=0}^{2m-1} \operatorname{sen} r x_j = 0.$$

Además, si r no es un múltiplo de m, estas sumas implican que

$$\sum_{j=0}^{2m-1} (\cos r x_j)^2 = \sum_{j=0}^{2m-1} \frac{1}{2}\left(1 + \cos 2r x_j\right) = \frac{1}{2}\left[2m + \sum_{j=0}^{2m-1} \cos 2r x_j\right] = \frac{1}{2}(2m + 0) = m$$

y, de igual forma, que

$$\sum_{j=0}^{2m-1} (\operatorname{sen} r x_j)^2 = \sum_{j=0}^{2m-1} \frac{1}{2}\left(1 - \cos 2r x_j\right) = m. \qquad \blacksquare$$

Ahora podemos mostrar la ortogonalidad establecida en la ecuación (8.24). Considere, por ejemplo, el caso

$$\sum_{j=0}^{2m-1} \phi_k(x_j)\phi_{n+l}(x_j) = \sum_{j=0}^{2m-1} (\cos k x_j)(\operatorname{sen} l x_j).$$

Puesto que

$$\cos k x_j \operatorname{sen} l x_j = \frac{1}{2}[\operatorname{sen}(l+k)x_j + \operatorname{sen}(l-k)x_j]$$

y tanto $(l + k)$ como $(l - k)$ son enteros que no son multiplicadores de $2m$, el lema 8.12 implica que

$$\sum_{j=0}^{2m-1} (\cos k x_j)(\operatorname{sen} l x_j) = \frac{1}{2}\left[\sum_{j=0}^{2m-1} \operatorname{sen}(l+k)x_j + \sum_{j=0}^{2m-1} \operatorname{sen}(l-k)x_j\right] = \frac{1}{2}(0+0) = 0.$$

Esta técnica se usa para mostrar que la condición de ortogonalidad se satisface para cualquier par de funciones y para producir el siguiente resultado.

Teorema 8.13 Las constantes en la suma

$$S_n(x) = \frac{a_0}{2} + a_n \cos nx + \sum_{k=1}^{n-1} (a_k \cos kx + b_k \operatorname{sen} kx)$$

que minimizan la suma de mínimos cuadrados

$$E(a_0, \ldots, a_n, b_1, \ldots, b_{n-1}) = \sum_{j=0}^{2m-1} (y_j - S_n(x_j))^2$$

son

- $a_k = \dfrac{1}{m} \displaystyle\sum_{j=0}^{2m-1} y_j \cos k x_j, \quad$ para cada $k = 0, 1, \ldots, n,$

y

- $b_k = \dfrac{1}{m} \displaystyle\sum_{j=0}^{2m-1} y_j \operatorname{sen} k x_j, \quad$ para cada $k = 1, 2, \ldots, n-1.$ ■

El teorema se prueba al establecer las derivadas parciales de E respecto a las a_k y las b_k para cero, como se realiza en las secciones 8.1 y 8.2 y al aplicar la ortogonalidad para simplificar las ecuaciones. Por ejemplo,

$$0 = \frac{\partial E}{\partial b_k} = 2 \sum_{j=0}^{2m-1} [y_j - S_n(x_j)](-\operatorname{sen} kx_j),$$

por lo que

$$\begin{aligned} 0 &= \sum_{j=0}^{2m-1} y_j \operatorname{sen} kx_j - \sum_{j=0}^{2m-1} S_n(x_j) \operatorname{sen} kx_j \\ &= \sum_{j=0}^{2m-1} y_j \operatorname{sen} kx_j - \frac{a_0}{2} \sum_{j=0}^{2m-1} \operatorname{sen} kx_j - a_n \sum_{j=0}^{2m-1} \operatorname{sen} kx_j \cos nx_j \\ &\quad - \sum_{l=1}^{n-1} a_l \sum_{j=0}^{2m-1} \operatorname{sen} kx_j \cos lx_j - \sum_{\substack{l=1,\\ l\neq k}}^{n-1} b_l \sum_{j=0}^{2m-1} \operatorname{sen} kx_j \operatorname{sen} lx_j - b_k \sum_{j=0}^{2m-1} (\operatorname{sen} kx_j)^2. \end{aligned}$$

La ortogonalidad implica que todas las sumas, excepto la primera y la última en el lado derecho son cero y el lema 8.12 establece que la suma final es m. Por lo tanto,

$$0 = \sum_{j=0}^{2m-1} y_j \operatorname{sen} kx_j - mb_k,$$

lo cual implica que

$$b_k = \frac{1}{m} \sum_{j=0}^{2m-1} y_j \operatorname{sen} kx_j.$$

El resultado para las a_k es similar, pero necesita un paso adicional para determinar a_0 (consulte el ejercicio 19).

Ejemplo 2 Encuentre $S_2(x)$, el polinomio trigonométrico de mínimos cuadrados discretos de grado 2 para $f(x) = 2x^2 - 9$ cuando x está en $[-\pi, \pi]$.

Solución Tenemos $m = 2(2) - 1 = 3$, por lo que los nodos son

$$x_j = \pi + \frac{j}{m}\pi \quad \text{y} \quad y_j = f(x_j) = 2x_j^2 - 9, \quad \text{para } j = 0, 1, 2, 3, 4, 5.$$

El polinomio trigonométrico es

$$S_2(x) = \frac{1}{2}a_0 + a_2 \cos 2x + (a_1 \cos x + b_1 \operatorname{sen} x),$$

donde

$$a_k = \frac{1}{3} \sum_{j=0}^{5} y_j \cos kx_j, \text{ para } k = 0, 1, 2, \quad \text{y} \quad b_1 = \frac{1}{3} \sum_{j=0}^{5} y_j \operatorname{sen} x_j.$$

Los coeficientes son

$$a_0 = \frac{1}{3}\left(f(-\pi) + f\left(-\frac{2\pi}{3}\right) + f\left(-\frac{\pi}{3}\right) + f(0) + f\left(\frac{\pi}{3}\right) + f\left(\frac{2\pi}{3}\right)\right)$$

$$= -4.10944566,$$

$$a_1 = \frac{1}{3}\left(f(-\pi)\cos(-\pi) + f\left(-\frac{2\pi}{3}\right)\cos\left(-\frac{2\pi}{3}\right) + f\left(-\frac{\pi}{3}\right)\cos\left(-\frac{\pi}{3}\right)\right.$$

$$\left. + f(0)\cos 0 + f\left(\frac{\pi}{3}\right)\cos\left(\frac{\pi}{3}\right) + f\left(\frac{2\pi}{3}\right)\cos\left(\frac{2\pi}{3}\right)\right) = -8.77298169,$$

$$a_2 = \frac{1}{3}\left(f(-\pi)\cos(-2\pi) + f\left(-\frac{2\pi}{3}\right)\cos\left(-\frac{4\pi}{3}\right) + f\left(-\frac{\pi}{3}\right)\cos\left(-\frac{2\pi}{3}\right)\right.$$

$$\left. + f(0)\cos 0 + f\left(\frac{\pi}{3}\right)\cos\left(\frac{2\pi}{3}\right) + f\left(\frac{2\pi}{3}\right)\cos\left(\frac{4\pi}{3}\right)\right) = 2.92432723,$$

y

$$b_1 = \frac{1}{3}\left(f(-\pi)\operatorname{sen}(-\pi) + f\left(-\frac{2\pi}{3}\right)\operatorname{sen}\left(-\frac{\pi}{3}\right) + f\left(-\frac{\pi}{3}\right)\left(-\frac{\pi}{3}\right)\right.$$

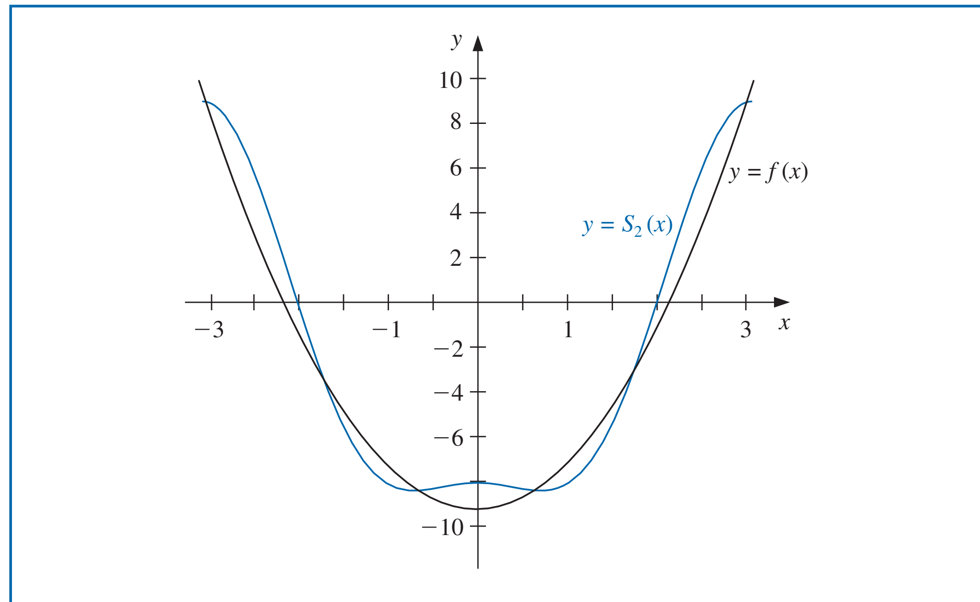
$$\left. + f(0)\operatorname{sen} 0 + f\left(\frac{\pi}{3}\right)\left(\frac{\pi}{3}\right) + f\left(\frac{2\pi}{3}\right)\left(\frac{2\pi}{3}\right)\right) = 0.$$

Por lo tanto,

$$S_2(x) = \frac{1}{2}(-4.10944562) - 8.77298169\cos x + 2.92432723\cos 2x.$$

La figura 8.5 muestra $f(x)$ y el polinomio trigonométrico de mínimos cuadrados discretos $S_2(x)$. ■

Figura 8.15



El siguiente ejemplo da una ilustración de cómo encontrar una aproximación por mínimos cuadrados para una función definida en un intervalo cerrado diferente a $[-\pi, \pi]$.

Ejemplo 3 Encuentre la aproximación por mínimos cuadrados discretos $S_3(x)$ para

$$f(x) = x^4 - 3x^3 + 2x^2 - \tan x(x-2) \text{ en } [0,2]$$

por medio de los datos $\{(x_j, y_j)\}_{j=0}^{9}$, donde $x_j = j/5$ y $y_j = f(x_j)$.

Solución Primero necesitamos la transformación lineal de [0, 2] a $[-\pi, \pi]$ dada por

$$z_j = \pi(x_j - 1).$$

Entonces, los datos transformados tienen la forma

$$\left\{\left(z_j, f\left(1 + \frac{z_j}{\pi}\right)\right)\right\}_{j=0}^{9}.$$

El polinomio trigonométrico de mínimos cuadrados es, por consiguiente,

$$S_3(z) = \left[\frac{a_0}{2} + a_3 \cos 3z + \sum_{k=1}^{2}(a_k \cos kz + b_k \operatorname{sen} kz)\right],$$

donde

$$a_k = \frac{1}{5}\sum_{j=0}^{9} f\left(1 + \frac{z_j}{\pi}\right)\cos kz_j, \quad \text{para } k = 0, 1, 2, 3,$$

y

$$b_k = \frac{1}{5}\sum_{j=0}^{9} f\left(1 + \frac{z_j}{\pi}\right)\operatorname{sen} kz_j, \quad \text{para } k = 1, 2.$$

La evaluación de estas sumas produce la aproximación

$$S_3(z) = 0.76201 + 0.77177\cos z + 0.017423\cos 2z + 0.0065673\cos 3z$$
$$- 0.38676 \operatorname{sen} z + 0.047806 \operatorname{sen} 2z,$$

y convertir de nuevo a las variables x nos da

$$S_3(x) = 0.76201 + 0.77177\cos \pi(x-1) + 0.017423\cos 2\pi(x-1)$$
$$+ 0.0065673\cos 3\pi(x-1) - 0.38676 \operatorname{sen} \pi(x-1) + 0.047806 \operatorname{sen} 2\pi(x-1).$$

La tabla 8.12 lista los valores de $f(x)$ y $S_3(x)$. ■

Tabla 8.12

x	$f(x)$	$S_3(x)$	$\|f(x) - S_3(x)\|$
0.125	0.26440	0.24060	2.38×10^{-2}
0.375	0.84081	0.85154	1.07×10^{-2}
0.625	1.36150	1.36248	9.74×10^{-4}
0.875	1.61282	1.60406	8.75×10^{-3}
1.125	1.36672	1.37566	8.94×10^{-3}
1.375	0.71697	0.71545	1.52×10^{-3}
1.625	0.07909	0.06929	9.80×10^{-3}
1.875	−0.14576	−0.12302	2.27×10^{-2}

La sección Conjunto de ejercicios 8.5 está disponible en línea. Encuentre la ruta de acceso en las páginas preliminares.

8.6 Transformadas rápidas de Fourier

En la última parte de la sección 8.5 determinamos la forma del polinomio de mínimos cuadrados discretos de grado n en los puntos de datos $2m - 1$ $\{(x_j, y_j)\}_{j=0}^{2m-1}$, donde $x_j = -\pi + (j/m)\pi$, para cada $j = 0, 1, \ldots, 2m - 1$.

El polinomio trigonométrico *de interpolación* en $\mathcal{T}_m$ en estos puntos de datos $2m$ es casi el mismo que el polinomio de mínimos cuadrados. Esto se debe a que el polinomio trigonométrico de mínimos cuadrados minimiza el término de error

$$E(S_m) = \sum_{j=0}^{2m-1} \left(y_j - S_m(x_j)\right)^2,$$

y para el polinomio trigonométrico de interpolación, este error es 0 y, por lo tanto, minimizado cuando $S_m(x_j) = y_j$, para cada $j = 0, 1, \ldots, 2m - 1$.

Se necesita una modificación de la forma del polinomio, sin embargo, si queremos que los coeficientes supongan la misma forma en el caso de mínimos cuadrados. En el lema 8.12, encontramos que si r no es un múltiplo de m, entonces

$$\sum_{j=0}^{2m-1} (\cos r x_j)^2 = m.$$

Por el contrario, la interpolación requiere calcular

$$\sum_{j=0}^{2m-1} (\cos m x_j)^2,$$

que (consulte el ejercicio 10) tiene el valor $2m$. Esto requiere que el polinomio de interpolación se escriba como

$$S_m(x) = \frac{a_0 + a_m \cos mx}{2} + \sum_{k=1}^{m-1} (a_k \cos kx + b_k \operatorname{sen} kx) \tag{8.26}$$

si queremos que la forma de las constantes a_k y b_k concuerde con las del polinomio de mínimos cuadrados discretos; es decir,

- $a_k = \dfrac{1}{m} \displaystyle\sum_{j=0}^{2m-1} y_j \cos kx_j, \quad$ para cada $k = 0, 1, \ldots, m$, y
- $b_k = \dfrac{1}{m} \displaystyle\sum_{j=0}^{2m-1} y_j \operatorname{sen} kx_j \quad$ para cada $k = 1, 2, \ldots, m - 1$.

La interpolación de grandes cantidades de datos espaciados uniformemente mediante polinomios trigonométricos puede producir resultados muy precisos. Es la técnica de aproximación adecuada en áreas relacionadas con filtros digitales, patrones de campo de antena, mecánica cuántica, óptica en numerosos problemas de simulación. Hasta mediados de la década de 1960, sin embargo, el método no se había aplicado ampliamente debido al número de cálculos aritméticos requeridos para determinar las constantes en la aproximación.

La interpolación de $2m$ puntos de datos mediante la técnica de cálculo directa requiere aproximadamente $(2m)^2$ multiplicaciones y $(2m)^2$ sumas. La aproximación de muchos miles de puntos de datos es muy común en áreas que requieren interpolación trigonométrica, por lo que los métodos directos para evaluar las constantes requieren operaciones de multiplicación y suma que llegan a millones. En general, el error de redondeo relacionado con este número de cálculos domina la aproximación.

En 1965, un artículo de J. W. Cooley y J. W. Turkey en el diario *Mathematics of Computation* [CT] (*Matemáticas de la computación*) describía un método diferente para calcular las constantes en el polinomio trigonométrico de interpolación. Este método sólo requiere $O(m \log_2 m)$ multiplicaciones y $O(m \log_2 m)$ sumas, siempre y cuando m sea seleccionada de manera adecuada. Para un problema con miles de puntos de datos, esto reduce el número de cálculos de millones a miles. En realidad, el método había sido descubierto algunos años antes de que apareciera el artículo de Cooley-Tukey pero había pasado desapercibido. ([Brigh], pp. 8–9, contiene un resumen histórico pero interesante del método).

El método descrito por Cooley y Tukey es conocido ya sea como **algoritmo Cooley-Tukey** o el **algoritmo de la transformada rápida de Fourier (FFT)** y ha conducido a una revolución en el uso de polinomios trigonométricos de interpolación. El método consiste en organizar el problema de tal forma que el número de puntos de datos utilizado se pueda factorizar fácilmente, de modo especial en potencias de dos.

En lugar de evaluar directamente las constantes a_k y b_k, el procedimiento de la transformada rápida de Fourier calcula los coeficientes c_k complejos en

$$\frac{1}{m}\sum_{k=0}^{2m-1} c_k e^{ikx}, \tag{8.27}$$

donde

$$c_k = \sum_{j=0}^{2m-1} y_j e^{ik\pi j/m}, \quad \text{para cada } k = 0, 1, \ldots, 2m-1. \tag{8.28}$$

Leonhard Euler proporcionó primero esta fórmula en 1748 en *Introductio in analysin infinitorum*, que hizo que las ideas de Johann Bernoulli fueran más precisas. Este trabajo basa los cálculos en la teoría de funciones elementales en lugar de curvas.

Una vez que las constantes c_k se han determinado, a_k y b_k se pueden recuperar por medio de la *fórmula de Euler*,

$$e^{iz} = \cos z + i \operatorname{sen} z.$$

Para cada $k = 0, 1, \ldots, m$, tenemos

$$\begin{aligned}
\frac{1}{m}c_k(-1)^k = \frac{1}{m}c_k e^{-i\pi k} &= \frac{1}{m}\sum_{j=0}^{2m-1} y_j e^{ik\pi j/m}e^{-i\pi k} = \frac{1}{m}\sum_{j=0}^{2m-1} y_j e^{ik(-\pi+(\pi j/m))}\\
&= \frac{1}{m}\sum_{j=0}^{2m-1} y_j\left(\cos k\left(-\pi+\frac{\pi j}{m}\right) + i \operatorname{sen} k\left(-\pi+\frac{\pi j}{m}\right)\right)\\
&= \frac{1}{m}\sum_{j=0}^{2m-1} y_j(\cos kx_j + i \operatorname{sen} kx_j).
\end{aligned}$$

Por lo que dada c_k, tenemos

$$a_k + ib_k = \frac{(-1)^k}{m}c_k. \tag{8.29}$$

Por conveniencia notacional, b_0 y b_m se suman al conjunto, pero ambos son 0 y no contribuyen a la suma resultante.

La característica de reducción-operación de la transformada rápida de Fourier resulta de calcular los coeficientes c_k en grupos y utiliza como relación básica el hecho de que para cualquier entero n,

$$e^{n\pi i} = \cos n\pi + i \operatorname{sen} n\pi = (-1)^n.$$

Suponga que $m = 2^p$ para algunos enteros positivos p. Para cada $k = 0, 1, \ldots, m - 1$, tenemos

$$c_k + c_{m+k} = \sum_{j=0}^{2m-1} y_j e^{ik\pi j/m} + \sum_{j=0}^{2m-1} y_j e^{i(m+k)\pi j/m} = \sum_{j=0}^{2m-1} y_j e^{ik\pi j/m}(1 + e^{\pi ij}).$$

Pero

$$1 + e^{i\pi j} = \begin{cases} 2, & \text{si } j \text{ es par}, \\ 0, & \text{si } j \text{ es impar}, \end{cases}$$

por lo que sólo hay que sumar m términos diferentes de cero.

Si j se reemplaza por $2j$ en el índice de la suma, podemos escribir la suma como

$$c_k + c_{m+k} = 2\sum_{j=0}^{m-1} y_{2j} e^{ik\pi(2j)/m};$$

es decir,

$$c_k + c_{m+k} = 2\sum_{j=0}^{m-1} y_{2j} e^{ik\pi j/(m/2)}. \tag{8.30}$$

De forma similar,

$$c_k - c_{m+k} = 2e^{ik\pi/m}\sum_{j=0}^{m-1} y_{2j+1} e^{ik\pi j/(m/2)}. \tag{8.31}$$

Puesto que c_k y c_{m+k} se pueden recuperar a partir de las ecuaciones (8.30) y (8.31), estas relaciones determinan todos los coeficientes c_k. También observe que las sumas en las ecuaciones (8.30) y (8.31) son de la misma forma que la suma en la ecuación (8.28), excepto que el índice m ha sido reemplazado por $m/2$.

Existen $2m$ coeficientes $c_0, c_1, \ldots, c_{2m-1}$ que deben calcularse. Usar la fórmula básica (8.28) requiere $2m$ multiplicaciones complejas por coeficiente, para un total de $(2m)^2$ operaciones. La ecuación (8.30) requiere m multiplicaciones complejas para cada $k = 0, 1, \ldots, m-1$, y la ecuación (8.31) requiere $m+1$ multiplicaciones complejas para cada $k = 0, 1, \ldots, m - 1$. Utilizar estas ecuaciones para calcular $c_0, c_1, \ldots, c_{2m-1}$ reducen el número de multiplicaciones complejas desde $(2m)^2 = 4m^2$ hasta

$$m \cdot m + m(m + 1) = 2m^2 + m.$$

Las sumas en las ecuaciones (8.30) y (8.31) tienen la misma forma que la original y m es una potencia de 2, por lo que la técnica de reducción se puede volver a aplicar a las sumas en las ecuaciones (8.30) y (8.31). Cada una de estas es reemplazada por dos sumas desde $j = 0$ hasta $j = (m/2) - 1$. Esto reduce la parte $2m^2$ de la suma a

$$2\left[\frac{m}{2} \cdot \frac{m}{2} + \frac{m}{2} \cdot \left(\frac{m}{2} + 1\right)\right] = m^2 + m.$$

Por lo que ahora se necesita un total de

$$(m^2 + m) + m = m^2 + 2m$$

multiplicaciones complejas en lugar de $(2m)^2$.

Al aplicar la técnica una vez más obtenemos cuatro sumas, cada una con $m/4$ términos y reduce la parte m^2 de este total a

$$4\left[\left(\frac{m}{4}\right)^2 + \frac{m}{4}\left(\frac{m}{4}+1\right)\right] = \frac{m^2}{2} + m,$$

para un total nuevo de $(m^2/2) + 3m$ multiplicaciones complejas. Repetir el proceso r veces reduce el número total de multiplicaciones complejas requeridas a

$$\frac{m^2}{2^{r-2}} + mr.$$

El proceso está completo cuando $r = p + 1$ porque entonces tenemos $m = 2^p$ y $2m = 2^{p+1}$. Por consiguiente, después de $r = p + 1$ reducciones de este tipo, el número de multiplicaciones complejas se reduce de $(2m)^2$ a

$$\frac{(2^p)^2}{2^{p-1}} + m(p+1) = 2m + pm + m = 3m + m\log_2 m = O(m\log_2 m).$$

Debido a la forma en que se ordenan los cálculos, el número de adiciones complejas requeridas es comparable.

Para ilustrar la significancia de esta reducción, suponga que tenemos $m = 2^{10} = 1\,024$. El cálculo directo de c_k, para $k = 0, 1, \ldots, 2m-1$, requeriría

$$(2m)^2 = (2048)^2 \approx 4\,200\,000$$

de cálculos. El procedimiento de la transformada rápida de Fourier reduce el número de cálculos a

$$3(1024) + 1024\log_2 1024 \approx 13\,300.$$

Ilustración Considere la técnica de la transformada rápida de Fourier aplicada a $8 = 2^3$ puntos de datos $\{(x_j, y_j)\}_{j=0}^{7}$, donde $x_j = -\pi + j\pi/4$, para cada $j = 0, 1, \ldots, 7$. En este caso, $2m = 8$, así $m = 4 = 2^2$ y $p = 2$.

A partir de la ecuación (8.26), tenemos

$$S_4(x) = \frac{a_0 + a_4\cos 4x}{2} + \sum_{k=1}^{3}(a_k\cos kx + b_k\,\text{sen}\,kx),$$

donde

$$a_k = \frac{1}{4}\sum_{j=0}^{7} y_j\cos kx_j \quad \text{y} \quad b_k = \frac{1}{4}\sum_{j=0}^{7} y_j\,\text{sen}\,kx_j, \quad k = 0, 1, 2, 3, 4.$$

Defina la transformada de Fourier como

$$\frac{1}{4}\sum_{j=0}^{7} c_k e^{ikx},$$

donde

$$c_k = \sum_{j=0}^{7} y_j e^{ik\pi j/4}, \quad \text{para } k = 0, 1, \ldots, 7.$$

Entonces mediante la ecuación (8.31) para $k = 0, 1, 2, 3, 4$, tenemos

$$\frac{1}{4}c_k e^{-ik\pi} = a_k + ib_k.$$

Mediante cálculo directo, las constantes complejas c_k están dadas por

$$c_0 = y_0 + y_1 + y_2 + y_3 + y_4 + y_5 + y_6 + y_7;$$

$$c_1 = y_0 + \left(\frac{i+1}{\sqrt{2}}\right) y_1 + iy_2 + \left(\frac{i-1}{\sqrt{2}}\right) y_3 - y_4 - \left(\frac{i+1}{\sqrt{2}}\right) y_5 - iy_6 - \left(\frac{i-1}{\sqrt{2}}\right) y_7;$$

$$c_2 = y_0 + iy_1 - y_2 - iy_3 + y_4 + iy_5 - y_6 - iy_7;$$

$$c_3 = y_0 + \left(\frac{i-1}{\sqrt{2}}\right) y_1 - iy_2 + \left(\frac{i+1}{\sqrt{2}}\right) y_3 - y_4 - \left(\frac{i-1}{\sqrt{2}}\right) y_5 + iy_6 - \left(\frac{i+1}{\sqrt{2}}\right) y_7;$$

$$c_4 = y_0 - y_1 + y_2 - y_3 + y_4 - y_5 + y_6 - y_7;$$

$$c_5 = y_0 - \left(\frac{i+1}{\sqrt{2}}\right) y_1 + iy_2 - \left(\frac{i-1}{\sqrt{2}}\right) y_3 - y_4 + \left(\frac{i+1}{\sqrt{2}}\right) y_5 - iy_6 + \left(\frac{i-1}{\sqrt{2}}\right) y_7;$$

$$c_6 = y_0 - iy_1 - y_2 + iy_3 + y_4 - iy_5 - y_6 + iy_7;$$

$$c_7 = y_0 - \left(\frac{i-1}{\sqrt{2}}\right) y_1 - iy_2 - \left(\frac{i+1}{\sqrt{2}}\right) y_3 - y_4 + \left(\frac{i-1}{\sqrt{2}}\right) y_5 + iy_6 + \left(\frac{i+1}{\sqrt{2}}\right) y_7.$$

Debido al pequeño tamaño del conjunto de puntos de datos, muchos de los coeficientes de y_j en estas ecuaciones son 1 o -1. Esta frecuencia disminuirá en una aplicación más grande, para contar las operaciones computacionales de manera precisa, la multiplicación por 1 o -1 se incluirá, a pesar de que no sería necesaria en este ejemplo. Con esta comprensión, se requieren 64 multiplicaciones/divisiones y 56 sumas/restas para el cálculo directo de $c_0, c_1, \ldots, c_7$.

Para aplicar el procedimiento de transformada rápida de Fourier con $r = 1$, primero determinamos

$$d_0 = \frac{c_0 + c_4}{2} = y_0 + y_2 + y_4 + y_6; \qquad d_4 = \frac{c_2 + c_6}{2} = y_0 - y_2 + y_4 - y_6;$$

$$d_1 = \frac{c_0 - c_4}{2} = y_1 + y_3 + y_5 + y_7; \qquad d_5 = \frac{c_2 - c_6}{2} = i(y_1 - y_3 + y_5 - y_7);$$

$$d_2 = \frac{c_1 + c_5}{2} = y_0 + iy_2 - y_4 - iy_6; \qquad d_6 = \frac{c_3 + c_7}{2} = y_0 - iy_2 - y_4 + iy_6;$$

$$d_3 = \frac{c_1 - c_5}{2} = \left(\frac{i+1}{\sqrt{2}}\right)(y_1 + iy_3 - y_5 - iy_7); \qquad d_7 = \frac{c_3 - c_7}{2} = \left(\frac{i-1}{\sqrt{2}}\right)(y_1 - iy_3 - y_5 + iy_7).$$

Entonces, definimos, para $r = 2$,

$$e_0 = \frac{d_0 + d_4}{2} = y_0 + y_4; \qquad e_4 = \frac{d_2 + d_6}{2} = y_0 - y_4;$$

$$e_1 = \frac{d_0 - d_4}{2} = y_2 + y_6; \qquad e_5 = \frac{d_2 - d_6}{2} = i(y_2 - y_6);$$

$$e_2 = \frac{id_1 + d_5}{2} = i(y_1 + y_5); \qquad e_6 = \frac{id_3 + d_7}{2} = \left(\frac{i-1}{\sqrt{2}}\right)(y_1 - y_5);$$

$$e_3 = \frac{id_1 - d_5}{2} = i(y_3 + y_7); \qquad e_7 = \frac{id_3 - d_7}{2} = i\left(\frac{i-1}{\sqrt{2}}\right)(y_3 - y_7).$$

Finalmente, para $r = p + 1 = 3$, definimos

$$f_0 = \frac{e_0 + e_4}{2} = y_0; \qquad f_4 = \frac{((i+1)/\sqrt{2})e_2 + e_6}{2} = \left(\frac{i-1}{\sqrt{2}}\right) y_1;$$

$$f_1 = \frac{e_0 - e_4}{2} = y_4; \qquad f_5 = \frac{((i+1)/\sqrt{2})e_2 - e_6}{2} = \left(\frac{i-1}{\sqrt{2}}\right) y_5;$$

$$f_2 = \frac{ie_1 + e_5}{2} = iy_2; \qquad f_6 = \frac{((i-1)/\sqrt{2})e_3 + e_7}{2} = \left(\frac{-i-1}{\sqrt{2}}\right) y_3;$$

$$f_3 = \frac{ie_1 - e_5}{2} = iy_6; \qquad f_7 = \frac{((i-1)/\sqrt{2})e_3 - e_7}{2} = \left(\frac{-i-1}{\sqrt{2}}\right) y_7.$$

$c_0, \ldots, c_7, d_0, \ldots, d_7, e_0, \ldots, e_7$, y $f_0, \ldots, f_7$ son independientes de los puntos de datos particulares; dependen solamente del hecho de que $m = 4$. Para cada m, existe un conjunto único de constantes $\{c_k\}_{k=0}^{2m-1}$, $\{d_k\}_{k=0}^{2m-1}$, $\{e_k\}_{k=0}^{2m-1}$, y $\{f_k\}_{k=0}^{2m-1}$. Esta parte del trabajo no es necesaria para una aplicación particular; sólo se requieren los siguientes cálculos:

f_k:

$$f_0 = y_0; \quad f_1 = y_4; \quad f_2 = iy_2; \quad f_3 = iy_6;$$

$$f_4 = \left(\frac{i-1}{\sqrt{2}}\right) y_1; \quad f_5 = \left(\frac{i-1}{\sqrt{2}}\right) y_5; \quad f_6 = -\left(\frac{i+1}{\sqrt{2}}\right) y_3; \quad f_7 = -\left(\frac{i+1}{\sqrt{2}}\right) y_7.$$

e_k:

$$e_0 = f_0 + f_1; \quad e_1 = -i(f_2 + f_3); \quad e_2 = -\left(\frac{i-1}{\sqrt{2}}\right)(f_4 + f_5);$$

$$e_3 = -\left(\frac{i+1}{\sqrt{2}}\right)(f_6 + f_7);$$

$$e_4 = f_0 - f_1; \quad e_5 = f_2 - f_3; \quad e_6 = f_4 - f_5; \quad e_7 = f_6 - f_7.$$

d_k:

$$d_0 = e_0 + e_1; \quad d_1 = -i(e_2 + e_3); \quad d_2 = e_4 + e_5; \quad d_3 = -i(e_6 + e_7);$$

$$d_4 = e_0 - e_1; \quad d_5 = e_2 - e_3; \quad d_6 = e_4 - e_5; \quad d_7 = e_6 - e_7.$$

c_k:

$$c_0 = d_0 + d_1; \quad c_1 = d_2 + d_3; \quad c_2 = d_4 + d_5; \quad c_3 = d_6 + d_7;$$

$$c_4 = d_0 - d_1; \quad c_5 = d_2 - d_3; \quad c_6 = d_4 - d_5; \quad c_7 = d_6 - d_7.$$

Calcular las constantes $c_0, c_1, \ldots, c_7$ de esta forma requiere el número de operaciones mostradas en la tabla 8.31. Observe de nuevo que la multiplicación por 1 o -1 se ha incluido en el conteo, a pesar de que no requiere esfuerzo computacional.

Tabla 8.13

Paso	Multiplicaciones/divisiones	Sumas/restas
(La f_k:)	8	0
(La e_k:)	8	8
(La d_k:)	8	8
(La c_k:)	0	8
Total	24	24

La falta de multiplicaciones/divisiones al encontrar c_k refleja el hecho de que para cualquier m, los coeficientes $\{c_k\}_{k=0}^{2m-1}$ se calculan a partir de $\{d_k\}_{k=0}^{2m-1}$ de la misma forma:

$$c_k = d_{2k} + d_{2k+1} \text{ y } c_{k+m} = d_{2k} - d_{2k+1}, \qquad \text{para } k = 0, 1, \ldots, m-1,$$

por lo que no existen multiplicaciones complejas.

En resumen, los cálculos directos de los coeficientes $c_0, c_1, \ldots, c_7$ requieren 64 multiplicaciones/divisiones y 56 sumas/restas. La técnica de la transformada rápida de Fourier reduce los cálculos a 24 multiplicaciones/divisiones y 24 sumas/restas. ■

El algoritmo 8.3 realiza la transformada rápida de Fourier cuando $m = 2^p$ para algunos enteros positivos p. Se pueden hacer modificaciones a la técnica cuando m toma otras formas.

ALGORITMO **8.3**

Transformada rápida de Fourier

Para calcular los coeficientes en la suma

$$\frac{1}{m}\sum_{k=0}^{2m-1} c_k e^{ikx} = \frac{1}{m}\sum_{k=0}^{2m-1} c_k(\cos kx + i \operatorname{sen} kx), \quad \text{donde } i = \sqrt{-1},$$

para los datos $\{(x_j, y_j)\}_{j=0}^{2m-1}$, donde $m = 2^p$ y $x_j = -\pi + j\pi/m$ para $j = 0, 1, \ldots, 2m-1$:

ENTRADA $m, p; y_0, y_1, \ldots, y_{2m-1}$.

SALIDA números complejos $c_0, \ldots, c_{2m-1}$; números reales $a_0, \ldots, a_m$; $b_1, \ldots, b_{m-1}$.

Paso 1 Determine $M = m$;
$q = p$;
$\zeta = e^{\pi i/m}$.

Paso 2 Para $j = 0, 1, \ldots, 2m-1$ determine $c_j = y_j$.

Paso 3 Para $j = 1, 2, \ldots, M$ determine $\xi_j = \zeta^j$;
$\xi_{j+M} = -\xi_j$.

Paso 4 Determine $K = 0$;
$\xi_0 = 1$.

Paso 5 Para $L = 1, 2, \ldots, p+1$ haga los pasos 6–12.

Paso 6 Mientras $K < 2m-1$ haga los pasos 7–11.

Paso 7 Para $j = 1, 2, \ldots, M$ haga los pasos 8–10.

Paso 8 Sea $K = k_p \cdot 2^p + k_{p-1} \cdot 2^{p-1} + \cdots + k_1 \cdot 2 + k_0$;
(*Descomponga* k.)
determine $K_1 = K/2^q = k_p \cdot 2^{p-q} + \cdots + k_{q+1} \cdot 2 + k_q$;
$K_2 = k_q \cdot 2^p + k_{q+1} \cdot 2^{p-1} + \cdots + k_p \cdot 2^q$.

Paso 9 Determine $\eta = c_{K+M}\xi_{K_2}$;
$c_{K+M} = c_K - \eta$;
$c_K = c_K + \eta$.

Paso 10 Determine $K = K + 1$.

Paso 11 Determine $K = K + M$.

Paso 12 Determine $K = 0$;
$M = M/2$;
$q = q - 1$.

Paso 13 Mientras $K < 2m - 1$ haga los pasos 14–16.

Paso 14 Sea $K = k_p \cdot 2^p + k_{p-1} \cdot 2^{p-1} + \cdots + k_1 \cdot 2 + k_0$; (*Decomponga k.*)
determine $j = k_0 \cdot 2^p + k_1 \cdot 2^{p-1} + \cdots + k_{p-1} \cdot 2 + k_p$.

Paso 15 Si $j > K$ entonces intercambie c_j y c_k.

Paso 16 Determine $K = K + 1$.

Paso 17 Determine $a_0 = c_0/m$;
$a_m = \text{Re}(e^{-i\pi m} c_m/m)$.

Paso 18 Para $j = 1, \ldots, m-1$ determine $a_j = \text{Re}(e^{-i\pi j} c_j/m)$;
$b_j = \text{Im}(e^{-i\pi j} c_j/m)$.

Paso 19 SALIDA $(c_0, \ldots, c_{2m-1}; a_0, \ldots, a_m; b_1, \ldots, b_{m-1})$;
PARE. ■

Ejemplo 1 Encuentre el polinomio trigonométrico de interpolación de grado 2 en $[-\pi, \pi]$ para los datos $\{(x_j, f(x_j))\}_{j=0}^{3}$, donde $f(x) = 2x^2 - 9$.

Solución Tenemos

$$a_k = \frac{1}{2} \sum_{j=0}^{3} f(x_j) \cos(kx_j) \quad \text{para } k = 0, 1, 2 \quad \text{y} \quad b_1 = \frac{1}{2} \sum_{j=0}^{3} f(x_j) \operatorname{sen}(x_j) \text{ por lo que,}$$

$$a_0 = \frac{1}{2}\left(f(-\pi) + f\left(-\frac{\pi}{2}\right) + f(0) + f\left(\frac{\pi}{2}\right)\right) = -3.19559339,$$

$$a_1 = \frac{1}{2}\left(f(-\pi)\cos(-\pi) + f\left(-\frac{\pi}{2}\right)\cos\left(-\frac{\pi}{2}\right) + f(0)\cos 0 + f\left(\frac{\pi}{2}\right)\cos\left(\frac{\pi}{2}\right)\right)$$
$$= -9.86960441,$$

$$a_2 = \frac{1}{2}\left(f(-\pi)\cos(-2\pi) + f\left(-\frac{\pi}{2}\right)\cos(-\pi) + f(0)\cos 0 + f\left(\frac{\pi}{2}\right)\cos(\pi)\right)$$
$$= 4.93480220,$$

y

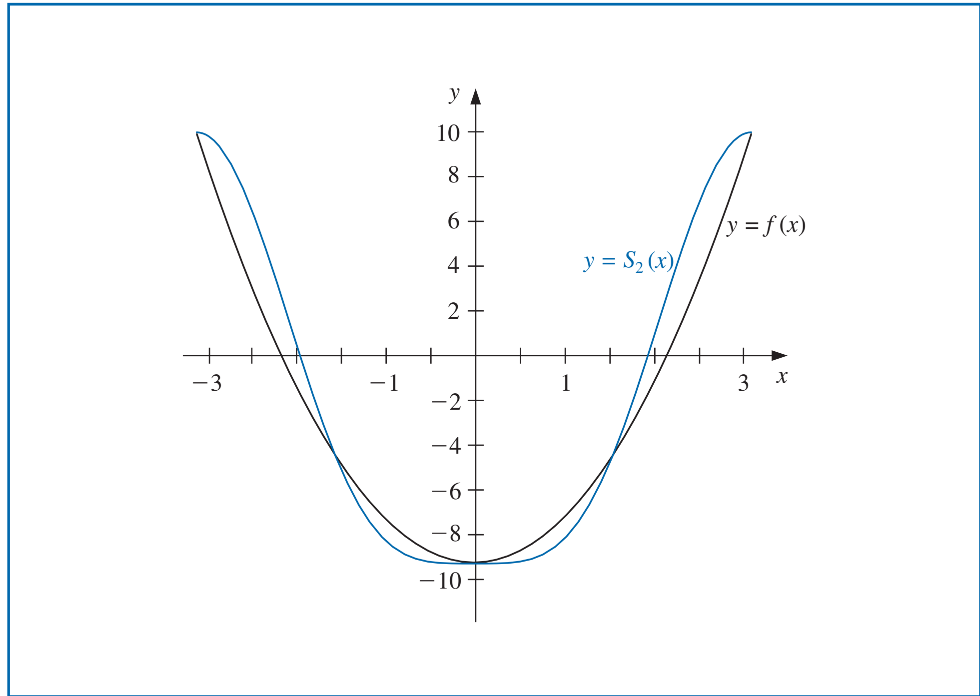
$$b_1 = \frac{1}{2}\left(f(-\pi)\operatorname{sen}(-\pi) + f\left(-\frac{\pi}{2}\right)\operatorname{sen}\left(-\frac{\pi}{2}\right) + f(0)\operatorname{sen} 0 + f\left(\frac{\pi}{2}\right)\operatorname{sen}\left(\frac{\pi}{2}\right)\right) = 0.$$

Por lo que,

$$S_2(x) = \frac{1}{2}(-3.19559339 + 4.93480220 \cos 2x) - 9.86960441 \cos x.$$

La figura 8.16 muestra $f(x)$ y el polinomio trigonométrico interpolante $S_2(x)$. ■

Figura 8.16



El siguiente ejemplo ilustra cómo encontrar un polinomio trigonométrico interpolante para una función definida en un intervalo cerrado diferente a $[-\pi, \pi]$.

Ejemplo 2 Determine el polinomio de interpolación trigonométrica de grado 4 en [0, 2] para los datos $\{(j/4,\ f(j/4))\}_{j=0}^{7}$, donde $f(x) = x^4 - 3x^3 + 2x^2 - \tan x(x-2)$.

Solución Primero necesitamos transformar el intervalo [0, 2] a $[-\pi, \pi]$. Esto está dado por

$$z_j = \pi(x_j - 1),$$

por lo que los datos de entrada para el algoritmo 8.3 son

$$\left\{z_j,\ f\left(1 + \frac{z_j}{\pi}\right)\right\}_{j=0}^{7}.$$

El polinomio de interpolación en z es

$$S_4(z) = 0.761979 + 0.771841 \cos z + 0.0173037 \cos 2z + 0.00686304 \cos 3z$$
$$- 0.000578545 \cos 4z - 0.386374 \operatorname{sen} z + 0.0468750 \operatorname{sen} 2z - 0.0113738 \operatorname{sen} 3z.$$

El polinomio trigonométrico $S_4(x)$ en [0, 2] se obtiene al sustituir $z = \pi(x-1)$ en $S_4(z)$. Las gráficas de $y = f(x)$ y $S_4(x)$ se muestran en la figura 8.17. Los valores de $f(x)$ y $S_4(x)$ están determinados en la tabla 8.14. ■

Figura 8.17

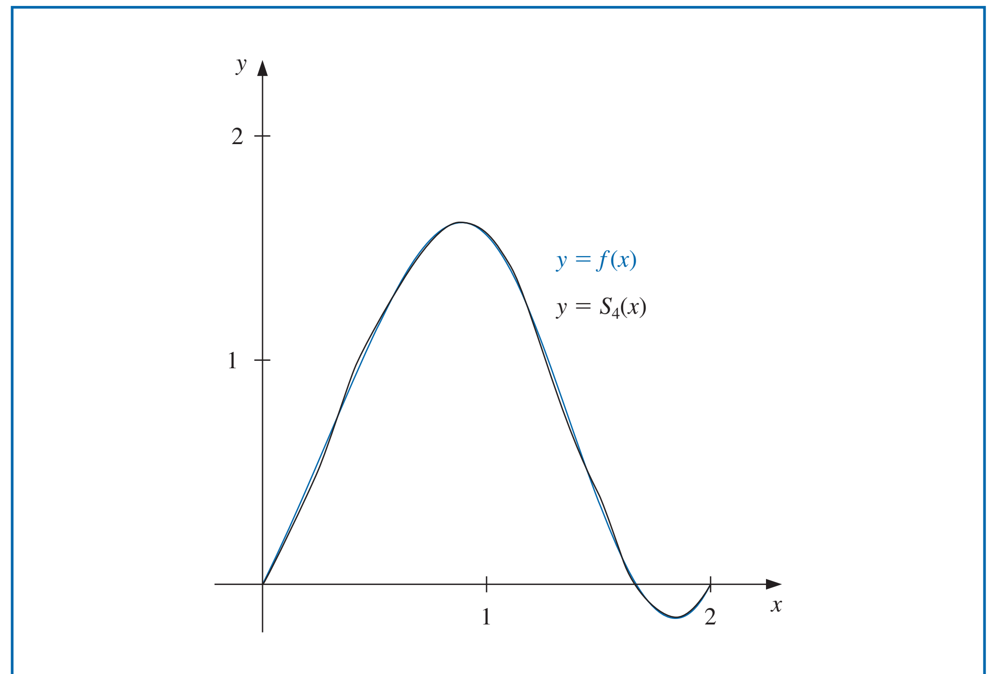


Tabla 8.14

x	$f(x)$	$S_4(x)$	$\lvert f(x) - S_4(x)\rvert$
0.125	0.26440	0.25001	1.44×10^{-2}
0.375	0.84081	0.84647	5.66×10^{-3}
0.625	1.36150	1.35824	3.27×10^{-3}
0.875	1.61282	1.61515	2.33×10^{-3}
1.125	1.36672	1.36471	2.02×10^{-3}
1.375	0.71697	0.71931	2.33×10^{-3}
1.625	0.07909	0.07496	4.14×10^{-3}
1.875	−0.14576	−0.13301	1.27×10^{-2}

Más detalles sobre la verificación de la validez del procedimiento de la transformada rápida de Fourier se pueden encontrar en [Ham], que presenta el método desde un enfoque matemático, o en [Brac], donde la presentación está basada en métodos que quizá sean familiares para los ingenieros.

[AHU], p. 252–269 es una buena referencia para un análisis de los aspectos computacionales del método. La modificación del procedimiento para el caso cuando m no es una potencia de 2 se puede encontrar en [Win]. Una presentación de las técnicas y el material relacionado desde el punto de vista del álgebra abstracta aplicada se da en [Lau], p. 438–465.

La sección Conjunto de ejercicios 8.6 está disponible en línea. Encuentre la ruta de acceso en las páginas preliminares.

8.7 Software numérico

La biblioteca IMSL proporciona un número de rutinas de aproximación incluyendo:

1. Los mínimos cuadrados lineales se ajustan a los datos con estadística
2. Los mínimos cuadrados discretos se ajustan a los datos con la selección del usuario de funciones de bases
3. Aproximación de mínimos cuadrados de spline cúbico

4. Aproximación racional ponderada de Chebyshev
5. La transformada rápida de Fourier se ajusta a los datos

La biblioteca NAG proporciona rutinas que incluyen el cálculo de lo siguiente:

1. La aproximación polinomial de mínimos cuadrados mediante una técnica para minimizar el error de redondeo
2. La aproximación de mínimos cuadrados de spline cúbico
3. Mejor ajuste en el sentido l_1
4. Mejor ajuste en el sentido l_∞
5. La transformada rápida de Fourier se ajusta a los datos

La biblioteca netlib contiene una rutina para calcular la aproximación de mínimos cuadrados polinomiales para un conjunto discreto de puntos y una rutina para evaluar este polinomio y cualquiera de sus derivadas en un punto determinado.

Las secciones Preguntas de análisis, Conceptos clave y Revisión del capítulo están disponibles en línea. Encuentre la ruta de acceso en las páginas preliminares.

CAPÍTULO

9 Aproximación de eigenvalores

Introducción

Las vibraciones longitudinales de un barra elástica de rigidez local $p(x)$ y densidad $\rho(x)$ se describen mediante la ecuación diferencial parcial

$$\rho(x)\frac{\partial^2 v}{\partial t^2}(x,t) = \frac{\partial}{\partial x}\left[p(x)\frac{\partial v}{\partial x}(x,t)\right],$$

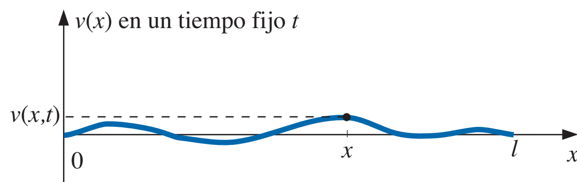
donde $v(x, t)$ es el desplazamiento longitudinal promedio de una sección de la barra desde su posición de equilibrio x en el tiempo t. Las vibraciones pueden escribirse como una suma de vibraciones armónicas simples:

$$v(x,t) = \sum_{k=0}^{\infty} c_k u_k(x)\cos\sqrt{\lambda_k}(t-t_0),$$

donde

$$\frac{d}{dx}\left[p(x)\frac{du_k}{dx}(x)\right] + \lambda_k\rho(x)u_k(x) = 0.$$

Si la barra tiene longitud l y está fija en sus extremos, entonces la ecuación diferencial se mantiene para $0 < x < l$ y $v(0) = v(l) = 0$.



Un sistema de estas ecuaciones diferenciales recibe el nombre de sistema Sturm-Liouville y los números λ_k son eigenvalores con eigenfunciones correspondientes $u_k(x)$.

Suponga que la barra es de 1 m de largo con rigidez uniforme $p(x) = p$ y densidad uniforme $\rho(x) = \rho$. Para aproximar u y λ, sea $h = 0.2$. Entonces $x_j = 0.2j$, para $0 \leq j \leq 5$,y podemos usar la fórmula de punto medio (4.5) en la sección 4.1 para aproximar las primeras derivadas. Esto da el sistema lineal

$$A\mathbf{w} = \begin{bmatrix} 2 & -1 & 0 & 0 \\ -1 & 2 & -1 & 0 \\ 0 & -1 & 2 & -1 \\ 0 & 0 & -1 & 2 \end{bmatrix}\begin{bmatrix} w_1 \\ w_2 \\ w_3 \\ w_4 \end{bmatrix} = -0.04\frac{\rho}{p}\lambda\begin{bmatrix} w_1 \\ w_2 \\ w_3 \\ w_4 \end{bmatrix} = -0.04\frac{\rho}{p}\lambda\mathbf{w}.$$

En este sistema, $w_j \approx u(x_j)$, para $1 \leq j \leq 4$, y $w_0 = w_5 = 0$. Los cuatro eigenvalores de A aproximan los eigenvalores del *sistema Sturm-Liouville*. Es la aproximación de los eigenvalores que consideraremos en este capítulo. En el ejercicio 7 de la sección 9.5 se analiza una aplicación Sturm-Liouville.

9.1 Álgebra lineal y eigenvalores

Los eigenvalores y los eigenvectores se presentaron en el capítulo 7 en relación con la convergencia de métodos iterativos para la aproximación de la solución de un sistema lineal. Para determinar los eigenvalores de una matriz A $n \times n$, construimos el polinomio característico

$$p(\lambda) = \det(A - \lambda I)$$

y, entonces, determinamos sus ceros. Encontrar el determinante de una matriz $n \times n$ es caro desde el punto computacional y hallar buenas aproximaciones para las raíces de $p(\lambda)$ también es difícil. En este capítulo exploraremos otros medios para aproximar los eigenvalores de una matriz. En la sección 9.6 damos una introducción a una técnica para la factorización de una matriz $m \times n$ en una forma que tiene aplicaciones valiosas en numerosas áreas.

En el capítulo 7 encontramos que una técnica iterativa para resolver un sistema lineal convergerá si todos los eigenvalores asociados con el problema tienen magnitud menor que 1. Los valores exactos de los eigenvalores en este caso no son muy importantes, sólo la región de un plano complejo en el que se encuentran. Un resultado importante en este sentido fue descubierto primero por S. A. Geršgorin. Es el tema de un libro muy interesante de Richard Varga [Var2].

Teorema 9.1 **(Círculo de Geršgorin)**

Semyon Aranovich Geršgorin (1901–1933) trabajó en el Instituto Tecnológico de Petrogrado hasta 1930, cuando se mudó al Instituto de Ingeniería Mecánica de Leningrado. Su artículo de 1931 *Über die Abgrenzung der Eigenwerte einer Matrix* ([Ger]) incluía lo que se conoce como teorema del círculo.

Sea A una matriz $n \times n$ y R_i denota el círculo en el plano complejo con centro a_{ii} y radio $\sum_{j=1,\, j\neq i}^{n} |a_{ij}|$; es decir,

$$R_i = \left\{ z \in \mathcal{C} \,\middle|\, |z - a_{ii}| \leq \sum_{j=1,\, j\neq i}^{n} |a_{ij}| \right\},$$

donde $\mathcal{C}$ denota el plano complejo. Los eigenvalores de A están contenidos en la unión de estos círculos, $R = \cup_{i=1}^{n} R_i$. Además, la unión de cualquier k de los círculos que no cruzan el resto de $(n - k)$, contiene precisamente k (multiplicidades contadas) de los eigenvalores.

Demostración Suponga que λ es un eigenvalor de A con un eigenvector asociado $\mathbf{x}$, donde $\|\mathbf{x}\|_\infty = 1$. Puesto que $A\mathbf{x} = \lambda\mathbf{x}$, la representación del componente equivalente es

$$\sum_{j=1}^{n} a_{ij}x_j = \lambda x_i, \quad \text{para cada } i = 1, 2, \ldots, n. \tag{9.1}$$

Sea k un entero con $|x_k| = \|\mathbf{x}\|_\infty = 1$. Cuando i $= k$, la ecuación (9.1) implica que

$$\sum_{j=1}^{n} a_{kj}x_j = \lambda x_k.$$

Por lo tanto,

$$\sum_{\substack{j=1,\\ j\neq k}}^{n} a_{kj}x_j = \lambda x_k - a_{kk}x_k = (\lambda - a_{kk})x_k,$$

y

$$|\lambda - a_{kk}| \cdot |x_k| = \left| \sum_{\substack{j=1,\\ j\neq k}}^{n} a_{kj} x_j \right| \leq \sum_{\substack{j=1,\\ j\neq k}}^{n} |a_{kj}||x_j|.$$

Pero, $|x_k| = \|\mathbf{x}\|_\infty = 1$, por lo que $|x_j| \leq |x_k| = 1$ para toda $j = 1, 2, \ldots, n$. Por lo tanto,

$$|\lambda - a_{kk}| \leq \sum_{\substack{j=1,\\ j\neq k}}^{n} |a_{kj}|.$$

Esto demuestra la primera afirmación en el teorema, que $\lambda \in R_k$. Una demostración se encuentra en [Var2], p. 8 o en [Or2], p. 48. ■

Ejemplo 1 Determine los círculos Gerŝgorin para la matriz

$$A = \begin{bmatrix} 4 & 1 & 1 \\ 0 & 2 & 1 \\ -2 & 0 & 9 \end{bmatrix}$$

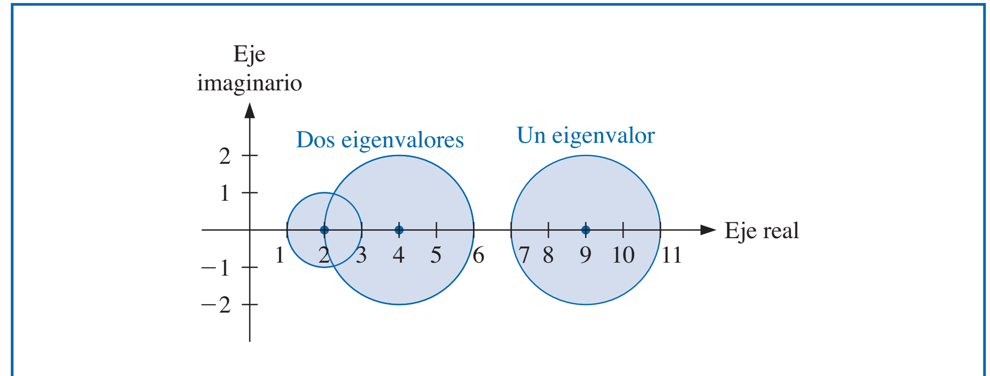
y úselos para encontrar los límites del radio espectral de A.

Solución Los círculos en el teorema de Gerŝgorin son (consulte la figura 9.1)

$$R_1 = \{z \in \mathcal{C} \mid |z-4| \leq 2\},\ \ R_2 = \{z \in \mathcal{C} \mid |z-2| \leq 1\},\ \text{ y } \ R_3 = \{z \in \mathcal{C} \mid |z-9| \leq 2\}.$$

Puesto que R_1 y R_2 están separados de R_3, existen precisamente dos eigenvalores dentro de $R_1 \cup R_2$ y uno dentro de R_3. Además, $\rho(A) = \text{máx}_{1\leq i\leq 3} |\lambda_i|$, por lo que $7 \leq \rho(A) \leq 11$. ■

Figura 9.1



Incluso cuando necesitamos encontrar los eigenvalores, muchas técnicas para su aproximación son iterativas. La determinación de las regiones en las que se encuentran es el primer paso para hallar las aproximaciones porque nos da aproximaciones iniciales.

Antes de considerar otros resultados concernientes a eigenvalores y a eigenvectores necesitamos algunas definiciones y resultados del álgebra lineal. Todos los resultados generales que se requerirán en lo que resta de este capítulo se listan aquí para facilitar la referencia su consulta. Las demostraciones de muchos de estos resultados no proporcionados se consideran en los ejercicios y es posible encontrar todos en diversos textos estándar sobre álgebra lineal (consulte, por ejemplo, [ND], [Poo] o [DG]).

La primera definición compara la definición de la independencia lineal de las funciones descritas en la sección 8.2. De hecho, mucho de lo que veremos en esta sección se compara con el material del capítulo 8.

Definición 9.2 Sea $\{\mathbf{v}^{(1)}, \mathbf{v}^{(2)}, \mathbf{v}^{(3)}, \dots, \mathbf{v}^{(k)}\}$ un conjunto de vectores. El conjunto es **linealmente independiente** si, siempre que

$$\mathbf{0} = \alpha_1 \mathbf{v}^{(1)} + \alpha_2 \mathbf{v}^{(2)} + \alpha_3 \mathbf{v}^{(3)} + \cdots + \alpha_k \mathbf{v}^{(k)},$$

entonces, $\alpha_i = 0$, para cada $i = 0, 1, \dots, k$. De lo contrario, el conjunto de vectores es **linealmente dependiente**. ■

Observe que cualquier conjunto de vectores que contienen el vector cero es linealmente dependiente.

Teorema 9.3 Suponga que $\{\mathbf{v}^{(1)}, \mathbf{v}^{(2)}, \mathbf{v}^{(3)}, \dots, \mathbf{v}^{(k)}\}$ es un conjunto de n vectores linealmente independientes en $\mathbb{R}^n$. Entonces, para cualquier vector $\mathbf{x} \in \mathbb{R}^n$, existe un único conjunto de constantes $\beta_1, \beta_2, \dots, \beta_n$ con

$$\mathbf{x} = \beta_1 \mathbf{v}^{(1)} + \beta_2 \mathbf{v}^{(2)} + \beta_3 \mathbf{v}^{(3)} + \cdots + \beta_n \mathbf{v}^{(n)}.$$

Demostración Sea A la matriz cuyas columnas son los vectores $\mathbf{v}^{(1)}, \mathbf{v}^{(2)}, \dots, \mathbf{v}^{(n)}$. Entonces, el conjunto $\{\mathbf{v}^{(1)}, \mathbf{v}^{(2)}, \dots, \mathbf{v}^{(n)}\}$ es linealmente independiente si y sólo si la ecuación matricial

$$A(\alpha_1, \alpha_2, \dots, \alpha_n)^t = \mathbf{0} \text{ tiene la única solución } (\alpha_1, \alpha_2, \dots, \alpha_n)^t = \mathbf{0}.$$

Pero, por el teorema 6.17 en la página 292, esto es equivalente a la ecuación matricial $A(\beta_1, \beta_2, \dots, \beta_n)^t = \mathbf{x}$, que tiene una única solución para cualquier vector $\mathbf{x} \in \mathbb{R}^n$. Esto, a su vez, es equivalente a la declaración de que para cualquier vector $\mathbf{x} \in \mathbb{R}^n$, existe un único conjunto de constantes $\beta_1, \beta_2, \dots, \beta_n$ con

$$\mathbf{x} = \beta_1 \mathbf{v}^{(1)} + \beta_2 \mathbf{v}^{(2)} + \beta_3 \mathbf{v}^{(3)} + \cdots + \beta_n \mathbf{v}^{(n)}.$$ ■

Definición 9.4 Cualquier conjunto de n vectores linealmente independientes en $\mathbb{R}^n$ recibe el nombre de **base** para $\mathbb{R}^n$. ■

Ejemplo 2 **a)** Muestre que $\mathbf{v}^{(1)} = (1, 0, 0)^t$, $\mathbf{v}^{(2)} = (-1, 1, 1)^t$, y $\mathbf{v}^{(3)} = (0, 4, 2)^t$ es una base para $\mathbb{R}^3$, y **b)** dado un vector arbitrario $\mathbf{x} \in \mathbb{R}^3$, encuentre β_1, β_2 y β_3 con

$$\mathbf{x} = \beta_1 \mathbf{v}^{(1)} + \beta_2 \mathbf{v}^{(2)} + \beta_3 \mathbf{v}^{(3)}.$$

Solución **a)** Sean α_1, α_2 y α_3 números con $\mathbf{0} = \alpha_1 \mathbf{v}^{(1)} + \alpha_2 \mathbf{v}^{(2)} + \alpha_3 \mathbf{v}^{(3)}$. Entonces

$$\begin{aligned}(0, 0, 0)^t &= \alpha_1 (1, 0, 0)^t + \alpha_2 (-1, 1, 1)^t + \alpha_3 (0, 4, 2)^t \\ &= (\alpha_1 - \alpha_2, \alpha_2 + 4\alpha_3, \alpha_2 + 2\alpha_3)^t,\end{aligned}$$

por lo que $\quad \alpha_1 - \alpha_2 = 0, \quad \alpha_2 + 4\alpha_3 = 0, \quad$ y $\quad \alpha_2 + 2\alpha_3 = 0.$

La única solución para este sistema es $\alpha_1 = \alpha_2 = \alpha_3 = 0$, por lo que este conjunto $\{\mathbf{v}^{(1)}, \mathbf{v}^{(2)}, \mathbf{v}^{(3)}\}$ de tres vectores linealmente independientes en $\mathbb{R}^3$ es una base para $\mathbb{R}^3$.

b) Sea $\mathbf{x} = (x_1, x_2, x_3)^t$ un vector en $\mathbb{R}^3$. Resolviendo

$$\begin{aligned}\mathbf{x} &= \beta_1 \mathbf{v}^{(1)} + \beta_2 \mathbf{v}^{(2)} + \beta_3 \mathbf{v}^{(3)} \\ &= \beta_1 (1, 0, 0)^t + \beta_2 (-1, 1, 1)^t + \beta_3 ((0, 4, 2)^t \\ &= (\beta_1 - \beta_2, \beta_2 + 4\beta_3, \beta_2 + 2\beta_3)^t\end{aligned}$$

es equivalente a resolver para β_1, β_2 y β_3 en el sistema

$$\beta_1 - \beta_2 = x_1, \quad \beta_2 + 4\beta_3 = x_2, \quad \beta_2 + 2\beta_3 = x_3.$$

Este sistema tiene la solución única

$$\beta_1 = x_1 - x_2 + 2x_3, \quad \beta_2 = 2x_3 - x_2 \text{ y } \beta_3 = \frac{1}{2}(x_2 - x_3).$$ ■

El siguiente resultado se usará en la sección 9.3 para desarrollar el método de potencia para aproximar los eigenvalores. En el ejercicio 12 se considera una prueba de este resultado.

Teorema 9.5 Si A es una matriz y $\lambda_1, \ldots, \lambda_k$ son eigenvalores distintos de A con eigenvectores asociados $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(k)}$, entonces $\{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(k)}\}$ es un conjunto linealmente independiente. ■

Ejemplo 3 Muestre que se puede formar una base para $\mathbb{R}^3$ usando los eigenvectores de la matriz 3×3

$$A = \begin{bmatrix} 2 & 0 & 0 \\ 1 & 1 & 2 \\ 1 & -1 & 4 \end{bmatrix}.$$

Solución En el ejemplo 2 de la sección 7.2 encontramos que A tiene el polinomio característico

$$p(\lambda) = p(A - \lambda I) = (\lambda - 3)(\lambda - 2)^2.$$

Por lo tanto, existen dos eigenvalores distintos de A: $\lambda_1 = 3$ y $\lambda_2 = 2$. En ese ejemplo también encontramos que $\lambda_1 = 3$ tiene el eigenvector $\mathbf{x}_1 = (0, 1, 1)^t$ y que hay dos eigenvectores linealmente independientes $\mathbf{x}_2 = (0, 2, 1)^t$ y $\mathbf{x}_3 = (-2, 0, 1)^t$ correspondientes a $\lambda_2 = 2$.

No es difícil mostrar (consulte el ejercicio 10) que este conjunto de tres eigenvectores

$$\{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3\} = \{(0, 1, 1)^t, (0, 2, 1)^t, (-2, 0, 1)^t\}$$

es linealmente independiente y, por lo tanto, forma una base para $\mathbb{R}^3$. ■

En el siguiente ejemplo observaremos una matriz cuyos eigenvalores son iguales a los del ejemplo 3, pero cuyos eigenvectores tienen un carácter diferente.

Ejemplo 4 Muestre que ningún conjunto de eigenvectores de la matriz 3×3

$$B = \begin{bmatrix} 2 & 1 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{bmatrix}$$

puede formar una base para $\mathbb{R}^3$.

Solución Esta matriz también tiene el mismo polinomio característico que la matriz A en el ejemplo 3:

$$p(\lambda) = \det \begin{bmatrix} 2-\lambda & 1 & 0 \\ 0 & 2-\lambda & 0 \\ 0 & 0 & 3-\lambda \end{bmatrix} = (\lambda - 3)(\lambda - 2)^2,$$

por lo que sus eigenvalores son iguales a los de A en el ejemplo 3, es decir, $\lambda_1 = 3$ y $\lambda_2 = 2$.

Para determinar los eigenvectores para B correspondientes al eigenvalor $\lambda_1 = 3$, necesitamos resolver el sistema $(B - 3I)\mathbf{x} = \mathbf{0}$, por lo que

$$\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} = (B - 3I)\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} -1 & 1 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 0 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} -x_1 + x_2 \\ -x_2 \\ 0, \end{bmatrix}.$$

Por lo tanto, $x_2 = 0$, $x_1 = x_2 = 0$ y x_3 es arbitrario. Haciendo $x_3 = 1$ esto nos da el único eigenvector linealmente independiente $(0, 0, 1)^t$ correspondiente a $\lambda_1 = 3$.

Considerando $\lambda_2 = 2$. Si

$$\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} = (B - 2\lambda) \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} x_2 \\ 0 \\ x_3, \end{bmatrix},$$

entonces $x_2 = 0, x_3 = 0$, y x_1 es arbitrario. Existe sólo un eigenvector linealmente independiente que corresponde a $\lambda_2 = 2$, lo que se puede expresar como $(1, 0, 0)^t$, aun cuando $\lambda_2 = 2$ fue un cero de multiplicidad dos del polinomio característico de B.

Es claro que estos dos eigenvectores no son suficientes para formar una base para $\mathbb{R}^3$. En particular, $(0, 1, 0)^t$ no es una combinación lineal de $\{(0, 0, 1)^t, (1, 0, 0)^t\}$. ■

Ahora veremos que cuando el número de eigenvectores linealmente independientes no corresponde al tamaño de la matriz, como en el caso del ejemplo 4, existen dificultades con los métodos de aproximación para encontrar los eigenvalores.

En la sección 8.2 consideramos los conjuntos ortogonales y ortonormales de funciones. Los vectores con estas propiedades se definen de forma similar.

Definición 9.6 Un conjunto de vectores $\{\mathbf{v}^{(1)}, \mathbf{v}^{(2)}, \ldots, \mathbf{v}^{(n)}\}$ recibe el nombre de **ortogonal** si $(\mathbf{v}^{(i)})^t\mathbf{v}^{(j)} = \mathbf{0}$, para toda $i \neq j$. Si, además $(\mathbf{v}^{(i)})^t\mathbf{v}^{(i)} = 1$, para toda $i = 1, 2, \ldots, n$. Entonces el conjunto recibe el nombre de **ortonormal**. ■

Puesto que $\mathbf{x}^t\mathbf{x} = \|\mathbf{x}\|_2^2$ para cualquier $\mathbf{x}$ en $\mathbb{R}^n$, un conjunto de vectores ortogonales $\{\mathbf{v}^{(1)}, \mathbf{v}^{(2)}, \ldots, \mathbf{v}^{(n)}\}$ es ortonormal si y sólo si

$$\|\mathbf{v}^{(i)}\|_2 = 1, \quad \text{para cada } i = 1, 2, \ldots, n.$$

Ejemplo 5 **a)** Muestre que los vectores $\mathbf{v}^{(1)} = (0, 4, 2)^t$, $\mathbf{v}^{(2)} = (-5, -1, 2)^t$ y $\mathbf{v}^{(3)} = (1, -1, 2)^t$ forman un conjunto ortogonal y **b)** úselos para determinar un conjunto de vectores ortonormales.

Solución **(a)** Tenemos $(\mathbf{v}^{(1)})^t\mathbf{v}^{(2)} = 0(-5) + 4(-1) + 2(2) = 0$,

$$(\mathbf{v}^{(1)})^t\mathbf{v}^{(3)} = 0(1) + 4(-1) + 2(2) = 0, \quad \text{y} \quad (\mathbf{v}^{(2)})^t\mathbf{v}^{(3)} = -5(1) - 1(-1) + 2(2) = 0,$$

por lo que los vectores son ortogonales y forman una base para $\mathbb{R}^n$. Las normas l_2 de estos vectores son

$$\|\mathbf{v}^{(1)}\|_2 = 2\sqrt{5}, \quad \|\mathbf{v}^{(2)}\|_2 = \sqrt{30}, \quad \text{y} \quad \|\mathbf{v}^{(3)}\|_2 = \sqrt{6}.$$

b) Los vectores

$$\mathbf{u}^{(1)} = \frac{\mathbf{v}^{(1)}}{\|\mathbf{v}^{(1)}\|_2} = \left(\frac{0}{2\sqrt{5}}, \frac{4}{2\sqrt{5}}, \frac{2}{2\sqrt{5}}\right)^t = \left(0, \frac{2\sqrt{5}}{5}, \frac{\sqrt{5}}{5}\right)^t,$$

$$\mathbf{u}^{(2)} = \frac{\mathbf{v}^{(2)}}{\|\mathbf{v}^{(2)}\|_2} = \left(\frac{-5}{\sqrt{30}}, \frac{-1}{\sqrt{30}}, \frac{2}{\sqrt{30}}\right)^t = \left(-\frac{\sqrt{30}}{6}, -\frac{\sqrt{30}}{30}, \frac{\sqrt{30}}{15}\right)^t, \text{ y}$$

$$\mathbf{u}^{(3)} = \frac{\mathbf{v}^{(3)}}{\|\mathbf{v}^{(3)}\|_2} = \left(\frac{1}{\sqrt{6}}, \frac{-1}{\sqrt{6}}, \frac{2}{\sqrt{6}}\right)^t = \left(\frac{\sqrt{6}}{6}, -\frac{\sqrt{6}}{6}, \frac{\sqrt{6}}{3}\right)^t$$

forman un conjunto ortonormal ya que heredan la ortogonalidad a partir de $\mathbf{v}^{(1)}, \mathbf{v}^{(2)}$, y $\mathbf{v}^{(3)}$. Además,

$$\|\mathbf{u}^{(1)}\|_2 = \|\mathbf{u}^{(2)}\|_2 = \|\mathbf{u}^{(3)}\|_2 = 1.$$ ■

La demostración del siguiente resultado se considera en el ejercicio 11.

Teorema 9.7 Un conjunto ortogonal de vectores diferentes a cero es linealmente independiente. ■

El proceso **Gram-Schmidt** para construir un conjunto de polinomios que son ortogonales respecto a la función de peso determinada se describió en el teorema 8.7 de la sección 8.2 (consulte la página 383). Existe un proceso paralelo, también conocido como Gram-Schmidt, que nos permite construir una base ortogonal para $\mathbb{R}^n$ dado un conjunto de n vectores linealmente independientes en $\mathbb{R}^n$.

Teorema 9.8 Sea $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k\}$ un conjunto de k vectores linealmente independientes en $\mathbb{R}^n$. Entonces $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k\}$ definido mediante

$$\mathbf{v}_1 = \mathbf{x}_1,$$

$$\mathbf{v}_2 = \mathbf{x}_2 - \left(\frac{\mathbf{v}_1^t\mathbf{x}_2}{\mathbf{v}_1^t\mathbf{v}_1}\right)\mathbf{v}_1,$$

$$\mathbf{v}_3 = \mathbf{x}_3 - \left(\frac{\mathbf{v}_1^t\mathbf{x}_3}{\mathbf{v}_1^t\mathbf{v}_1}\right)\mathbf{v}_1 - \left(\frac{\mathbf{v}_2^t\mathbf{x}_3}{\mathbf{v}_2^t\mathbf{v}_2}\right)\mathbf{v}_2,$$

$$\vdots$$

$$\mathbf{v}_k = \mathbf{x}_k - \sum_{i=1}^{k-1}\left(\frac{\mathbf{v}_i^t\mathbf{x}_k}{\mathbf{v}_i^t\mathbf{v}_i}\right)\mathbf{v}_i$$

es un conjunto de k vectores ortogonales en $\mathbb{R}^n$. ■

La demostración de este teorema, que se analiza en el ejercicio 16, es una verificación directa del hecho de que para cada $1 \le i \le k$ y $1 \le j \le k$ y con $i \ne j$, tenemos $\mathbf{v}_i^t\mathbf{v}_j = 0$.

Observe que cuando el conjunto original de vectores forma una base para $\mathbb{R}^n$, es decir, cuando $k = n$, entonces los vectores construidos forman una base ortogonal para $\mathbb{R}^n$. A partir de esto podemos formar una base ortonormal $\{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n\}$ simplemente al definir para cada $i = 1, 2, \ldots, n$

$$\mathbf{u}_i = \frac{\mathbf{v}_i}{||\mathbf{v}_i||_2}.$$

El siguiente ejemplo ilustra cómo se puede construir una base ortogonal para $\mathbb{R}^3$ a partir de tres vectores linealmente independientes en $\mathbb{R}^3$.

Ejemplo 6 Use el proceso Gram-Schmidt para determinar un conjunto de vectores ortogonales a partir de los vectores linealmente independientes

$$\mathbf{x}^{(1)} = (1, 0, 0)^t, \quad \mathbf{x}^{(2)} = (1, 1, 0)^t, \quad \text{y} \quad \mathbf{x}^{(3)} = (1, 1, 1)^t.$$

Solución Tenemos los vectores ortogonales $\mathbf{v}^{(1)}$, $\mathbf{v}^{(2)}$ y $\mathbf{v}^{(3)}$, dados por

$$\mathbf{v}^{(1)} = \mathbf{x}^{(1)} = (1, 0, 0)^t$$

$$\mathbf{v}^{(2)} = (1, 1, 0)^t - \left(\frac{((1, 0, 0)^t)^t(1, 1, 0)^t}{((1, 0, 0)^t)^t(1, 0, 0)^t}\right)(1, 0, 0)^t = (1, 1, 0)^t - (1, 0, 0)^t = (0, 1, 0)^t$$

$$\mathbf{v}^{(3)} = (1, 1, 1)^t - \left(\frac{((1, 0, 0)^t)^t(1, 1, 1)^t}{((1, 0, 0)^t)^t(1, 0, 0)^t}\right)(1, 0, 0)^t - \left(\frac{((0, 1, 0)^t)^t(1, 1, 1)^t}{((0, 1, 0)^t)^t(0, 1, 0)^t}\right)(0, 1, 0)^t$$

$$= (1, 1, 1)^t - (1, 0, 0)^t - (0, 1, 0)^t = (0, 0, 1)^t.$$

El conjunto $\{\mathbf{v}^{(1)}, \mathbf{v}^{(2)}, \mathbf{v}^{(3)}\}$ resulta ser tanto ortonormal, como ortogonal, pero comúnmente, ésta no es la situación. ■

La sección Conjunto de ejercicios 9.1 está disponible en línea. Encuentre la ruta de acceso en las páginas preliminares.

9.2 Matrices ortogonales y transformaciones de similitud

En esta sección consideraremos la conexión entre los conjuntos de vectores y las matrices formadas usando estos vectores como sus columnas. Primero consideramos algunos resultados alrededor de una clase de matrices especiales. La terminología en la siguiente definición sigue del hecho de que las columnas de una matriz ortogonal formarán un conjunto ortogonal de vectores.

Definición 9.9 Se dice que una matriz Q es **ortogonal** si sus columnas $\{\mathbf{q}_1^t, \mathbf{q}_2^t, \ldots, \mathbf{q}_n^t\}$ forman un conjunto ortonormal en $\mathbb{R}^n$. ■

Probablemente sería mejor llamar *ortonormales* a las matrices ortogonales porque las columnas no sólo forman un conjunto ortogonal de vectores sino también uno ortonormal.

Las siguientes propiedades importantes de las matrices ortogonales se consideran en el ejercicio 19.

Teorema 9.10 Suponga que Q es una matriz $n \times n$ ortogonal. Entonces

i) Q es invertible con $Q^{-1} = Q^t$.

ii) Para cualquier $\mathbf{x}$ y $\mathbf{y}$ en $\mathbb{R}^n$, $(Q\mathbf{x})^t Q\mathbf{y} = \mathbf{x}^t\mathbf{y}$.

iii) Para cualquier $\mathbf{x}$ en $\mathbb{R}^n$, $||Q\mathbf{x}||_2 = ||\mathbf{x}||_2$.

iv) Cualquier matriz invertible Q con $Q^{-1} = Q^t$ es ortogonal. ■

Como ejemplo, las matrices de permutación que se han analizado en la sección 6.5 tienen esta propiedad, por lo que son ortogonales.

A menudo, la propiedad **iii)** del teorema 9.10 se expresa al establecer que las matrices ortogonales preservan la norma l_2. Como consecuencia inmediata de esta propiedad, todas las matrices ortogonales Q tienen $||Q||_2 = 1$.

Ejemplo 1 Muestre que la matriz

$$Q = [\mathbf{u}^{(1)}, \mathbf{u}^{(2)}, \mathbf{u}^{(3)}] = \begin{bmatrix} 0 & -\frac{\sqrt{30}}{6} & \frac{\sqrt{6}}{6} \\ \frac{2\sqrt{5}}{5} & -\frac{\sqrt{30}}{30} & -\frac{\sqrt{6}}{6} \\ \frac{\sqrt{5}}{5} & \frac{\sqrt{30}}{15} & \frac{\sqrt{6}}{3} \end{bmatrix}$$

formada a partir del conjunto ortonormal de vectores encontrado en el ejemplo 5 de la sección 9.1 es una matriz ortogonal.

Solución Observe que

$$QQ^t = \begin{bmatrix} 0 & \frac{-\sqrt{30}}{6} & \frac{\sqrt{6}}{6} \\ \frac{2\sqrt{5}}{5} & -\frac{\sqrt{30}}{30} & -\frac{\sqrt{6}}{6} \\ \frac{\sqrt{5}}{5} & \frac{\sqrt{30}}{15} & \frac{\sqrt{6}}{3} \end{bmatrix} \cdot \begin{bmatrix} 0 & \frac{2\sqrt{5}}{5} & \frac{\sqrt{5}}{5} \\ -\frac{\sqrt{30}}{6} & -\frac{\sqrt{30}}{30} & \frac{\sqrt{30}}{15} \\ \frac{\sqrt{6}}{6} & -\frac{\sqrt{6}}{6} & \frac{\sqrt{6}}{3} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = I.$$

Mediante el corolario 6.18 en la sección 6.4 (consulte la página 298), ésto es suficiente para garantizar que $Q^t = Q^{-1}$. Por ello, Q es una matriz ortogonal. ■

La siguiente definición establece las bases de muchas técnicas para determinar los eigenvalores de una matriz.

Definición 9.11 Se dice que dos matrices A y B son **similares** si existe una matriz no singular S con $A = S^{-1}BS$. ■

Una característica importante de las matrices similares es que tienen los mismos eigenvalores.

Teorema 9.12 Suponga que A y B son matrices similares con $A = S^{-1}BS$ y λ es un eigenvalor de A con un eigenvector $\mathbf{x}$ relacionado. Entonces λ es un eigenvalor de B con un eigenvector relacionado $S\mathbf{x}$.

Demostración Sea $\mathbf{x} \neq \mathbf{0}$ tal que

$$S^{-1}BS\mathbf{x} = A\mathbf{x} = \lambda\mathbf{x}.$$

Multiplicando a la izquierda por la matriz S da

$$BS\mathbf{x} = \lambda S\mathbf{x}.$$

Puesto que $\mathbf{x} \neq \mathbf{0}$ y S es no singular, $S\mathbf{x} \neq \mathbf{0}$. Por lo tanto, $S\mathbf{x}$ es un eigenvector de $\boldsymbol{B}$ correspondiente a su eigenvalor λ. ■

Un uso especialmente importante de similitud se presenta cuando la matriz A $n \times n$ es similar a la matriz diagonal, es decir, cuando existe una matriz diagonal D y una matriz invertible S con

$$A = S^{-1}DS \text{ o, de manera equivalente, } D = SAS^{-1}.$$

El siguiente resultado no es difícil de mostrar. Se considera en el ejercicio 20

Teorema 9.13 Una matriz A $n \times n$ es similar a una matriz diagonal D si y sólo si A tiene n eigenvectores linealmente independientes. En este caso, $D = S^{-1}AS$, donde las columnas de S consisten en eigenvectores y el i-ésimo elemento diagonal de D es el eigenvalor de A que corresponde a la i-ésima columna de S. ■

El par de matrices S y D no es único. Por ejemplo, cualquier reordenamiento de las columnas de S y el reordenamiento correspondiente de los elementos de la diagonal de D darán un par distinto. Consulte el ejercicio 13 para una ilustración.

En el teorema 9.5 observamos que los eigenvectores de una matriz que corresponden a los distintos eigenvalores forman un conjunto linealmente independiente. Como consecuencia, tenemos el siguiente corolario para el teorema 9.13.

Corolario 9.14 Una matriz A $n \times n$ que tiene n eigenvalores diferentes es similar a una matriz diagonal. ■

De hecho, no necesitamos que la matriz de similitud sea diagonal para que este concepto sea útil. Suponga que A es similar a la matriz triangular B. La determinación de los eigenvalores es fácil para una matriz triangular B, porque en este caso λ es una solución para la ecuación

$$0 = \det(B - \lambda I) = \prod_{i=1}^{n}(b_{ii} - \lambda)$$

si y sólo si $\lambda = b_{ii}$ para algunas i. El siguiente resultado describe una relación, llamada **transformación de similitud**, entre las matrices arbitrarias y las triangulares.

Teorema 9.15 **(Teorema de Schur)**

Sea A una matriz arbitraria. Existe una matriz no singular U con la propiedad de que

$$T = U^{-1}AU,$$

donde T es una matriz triangular superior, cuyas entradas diagonales consisten en eigenvalores de A. ■

Issai Schur (1875–1941) es conocido principalmente por su trabajo en la teoría de grupos, pero también trabajó en la teoría de números, el análisis y otras áreas. Publicó lo que se conoce como teorema de Schur en 1909.

La norma l_2 de una matriz unitaria es *1*.

La matriz U, cuya existencia está garantizada por el teorema 9.15, satisface la condición $\|U\mathbf{x}\|_2 = \|\mathbf{x}\|_2$ para cualquier vector $\mathbf{x}$. Las matrices con esta propiedad reciben el nombre de **unitarias**. A pesar de que no usaremos esta propiedad de preservación de la norma, sí aumenta significativamente la aplicación del teorema de Schur.

El teorema 9.15 es un teorema de existencia que garantiza que existe la matriz T, pero no proporciona un medio constructivo para encontrar T ya que requiere un conocimiento de los eigenvalores de A. En muchos casos, es demasiado difícil determinar la transformación de similitud U.

El siguiente resultado para las matrices simétricas reduce la complicación porque, en este caso, la matriz de transformación es ortogonal.

Teorema 9.16 La matriz A $n \times n$ es simétrica si y sólo si existe una matriz diagonal D y una matriz ortogonal Q con $A = QDQ^t$.

Demostración Primero suponga que $A = QDQ^t$, donde Q es ortogonal y D es diagonal. Entonces

$$A^t = \left(QDQ^t\right)^t = \left(Q^t\right)^t D\,Q^t = QDQ^t = A,$$

y A es simétrica.

Para demostrar que todas las matrices simétricas A se pueden escribir de la forma $A = QDQ^t$, primero considere los diferentes eigenvalores de A. Si $A\mathbf{v}_1 = \lambda_1\mathbf{v}_1$ y $A\mathbf{v}_2 = \lambda_2\mathbf{v}_2$, con $\lambda_1 \neq \lambda_2$, entonces, puesto que $A^t = A$, tenemos

$$(\lambda_1 - \lambda_2)\mathbf{v}_1^t\mathbf{v}_2 = (\lambda_1\mathbf{v}_1)^t\mathbf{v}_2 - \mathbf{v}_1^t(\lambda_2\mathbf{v}_2) = (A\mathbf{v}_1)^t\mathbf{v}_2 - \mathbf{v}_1^t(A\mathbf{v}_2) = \mathbf{v}_1^t A^t\mathbf{v}_2 - \mathbf{v}_1^t A\mathbf{v}_2 = 0,$$

por lo que $\mathbf{v}_1^t\mathbf{v}_2 = 0$. Por lo tanto, seleccionamos vectores ortonormales para diferentes eigenvalores simplemente al normalizar todos estos eigenvectores ortogonales. Cuando los eigenvalores no son distintos, habrá subespacios de eigenvectores para cada uno de los múltiples eigenvalores y con la ayuda del proceso de ortogonalización de Gram-Schmidt, podemos encontrar un conjunto completo de n eigenvectores ortonormales. ■

El siguiente corolario para el teorema 9.16 demuestra algunas de las propiedades más interesantes de las matrices simétricas.

Corolario 9.17 Suponga que A es una matriz simétrica $n \times n$. Entonces existen n eigenvectores de A que forman un conjunto ortonormal y los eigenvalores de A son números reales. ■

Demostración Si $Q = (q_{ij})$ y $D = (d_{ij})$ son las matrices especificadas en el teorema 9.16, entonces

$$D = Q^tAQ = Q^{-1}AQ \quad \text{implica que} \quad AQ = QD.$$

Sea $1 \le i \le n$ y $\mathbf{v}_i = (q_{1i}, q_{2i}, \ldots, q_{ni})^t$ la i-ésima columna de Q. Entonces

$$A\mathbf{v}_i = d_{ii}\mathbf{v}_i,$$

y d_{ii} es un eigenvalor de A con eigenvector $\mathbf{v}_i$, la i-ésima columna de Q. Las columnas de Q son ortonormales, por lo que los eigenvectores de A son ortonormales.

Al multiplicar esta ecuación a la izquierda por $\mathbf{v}_i^t$ obtenemos

$$\mathbf{v}_i^tA\mathbf{v}_i = d_{ii}\mathbf{v}_i^t\mathbf{v}_i.$$

Puesto que $\mathbf{v}_i^t A\mathbf{v}_i$ y $\mathbf{v}_i^t\mathbf{v}_i$ son números reales y $\mathbf{v}_i^t\mathbf{v}_i = 1$, el eigenvalor $d_{ii} = \mathbf{v}_i^t A\mathbf{v}_i$ es un número real, para cada $i = 1, 2, \ldots, n$. ■

A veces, una matriz simétrica cuyos eigenvalores son todos números reales no negativos recibe el nombre de *definida no negativa* (o *semidefinida positiva*).

Recuerde de la sección 6.6 que una matriz simétrica A es llamada definida positiva si para todos los vectores diferentes de cero, se tiene $\mathbf{x}^t A\mathbf{x} > 0$. El siguiente teorema caracteriza a las matrices definidas positivas en términos de eigenvalores. Esta propiedad de eigenvalor hace que las matrices definidas positivas sean importantes en las aplicaciones.

Teorema 9.18 Una matriz simétrica A es definida positiva si y sólo si todos los eigenvalores de A son positivos.

Demostración Primero suponga que A es definida positiva y que λ es un eigenvalor de A con un eigenvector asociado $\mathbf{x}$, con $||\mathbf{x}||_2 = 1$. Entonces

$$0 < \mathbf{x}^t A\mathbf{x} = \lambda\mathbf{x}^t\mathbf{x} = \lambda\|\mathbf{x}\|_2^2 = \lambda.$$

Para mostrar el recíproco, suponga que A es simétrica con eigenvalores positivos. Por el corolario 9.17, A tiene n eigenvectores, $\mathbf{v}^{(1)}, \mathbf{v}^{(2)}, \ldots, \mathbf{v}^{(n)}$, que forman un conjunto ortonormal y, por el teorema 9.7, un conjunto linealmente independiente. Por lo tanto, para cualquier $\mathbf{x} \neq \mathbf{0}$, existe un único conjunto de constantes diferentes de cero $\beta_1, \beta_2, \ldots, \beta_n$ para las que

$$\mathbf{x} = \sum_{i=1}^{n} \beta_i \mathbf{v}^{(i)}.$$

Al multiplicar por $\mathbf{x}^t A$ obtenemos

$$\mathbf{x}^t A\mathbf{x} = \mathbf{x}^t \left(\sum_{i=1}^{n} \beta_i A\mathbf{v}^{(i)}\right) = \mathbf{x}^t \left(\sum_{i=1}^{n} \beta_i \lambda_i \mathbf{v}^{(i)}\right) = \sum_{j=1}^{n}\sum_{i=1}^{n} \beta_j\beta_i\lambda_i (\mathbf{v}^{(j)})^t \mathbf{v}^{(i)}.$$

Pero los vectores $\mathbf{v}^{(1)}, \mathbf{v}^{(2)}, \ldots, \mathbf{v}^{(n)}$ forman un conjunto ortonormal, por lo que

$$(\mathbf{v}^{(j)})^t \mathbf{v}^{(i)} = \begin{cases} 0, & \text{si } i \neq j, \\ 1, & \text{si } i = j. \end{cases}$$

Esto, junto con el hecho de que λ_i son todas positivas, implica que

$$\mathbf{x}^t A\mathbf{x} = \sum_{j=1}^{n}\sum_{i=1}^{n} \beta_j\beta_i\lambda_i (\mathbf{v}^{(j)})^t \mathbf{v}^{(i)} = \sum_{i=1}^{n} \lambda_i \beta_i^2 > 0.$$

Por lo tanto, A es definida positiva. ■

La sección Conjunto de ejercicios 9.2 está disponible en línea. Encuentre la ruta de acceso en las páginas preliminares.

9.3 El método de potencia

El **método de potencia** es una técnica iterativa que se usa para determinar el eigenvalor dominante de una matriz (es decir, el eigenvalor con la mayor magnitud). Al modificar ligeramente el método, también se puede usar para determinar otros eigenvalores. Una característica útil del método de potencia es que no sólo produce un eigenvalor, sino también un eigenvector asociado. De hecho, a menudo, el método de potencia se aplica para encontrar un eigenvector para un eigenvalor que es determinado por algunos otros medios.

El nombre de método de potencia se deriva del hecho de que las iteraciones exageran el tamaño relativo de las magnitudes de los eigenvalores.

Para aplicar el método de potencia, suponemos que la matriz $n \times n$ A tiene n eigenvalores $\lambda_1, \lambda_2, \ldots, \lambda_n$, con un conjunto asociado de eigenvectores linealmente independientes $\{\mathbf{v}^{(1)}, \mathbf{v}^{(2)}, \mathbf{v}^{(3)}, \ldots, \mathbf{v}^{(n)}\}$. Además, suponemos que A tiene exactamente un eigenvalor λ_1, que es más grande en magnitud, por lo que

$$|\lambda_1| > |\lambda_2| \geq |\lambda_3| \geq \cdots \geq |\lambda_n| \geq 0.$$

El ejemplo 4 de la sección 9.1 ilustra que una matriz $n \times n$ no necesita tener n eigenvectores linealmente independientes. Cuando esto no es así, el método de potencia puede seguir siendo exitoso, pero no se garantiza que lo sea.

Si $\mathbf{x}$ es cualquier vector en $\mathbb{R}^n$, el hecho de que $\{\mathbf{v}^{(1)}, \mathbf{v}^{(2)}, \mathbf{v}^{(3)}, \ldots, \mathbf{v}^{(n)}\}$ es linealmente independiente implica que existen constantes $\beta_1, \beta_2, \ldots, \beta_n$ con

$$\mathbf{x} = \sum_{j=1}^{n} \beta_j \mathbf{v}^{(j)}.$$

Al multiplicar ambos lados de esta ecuación por $A, A^2, \ldots, A^k, \ldots$ obtenemos

$$A\mathbf{x} = \sum_{j=1}^{n} \beta_j A\mathbf{v}^{(j)} = \sum_{j=1}^{n} \beta_j \lambda_j \mathbf{v}^{(j)}, \quad A^2\mathbf{x} = \sum_{j=1}^{n} \beta_j \lambda_j A\mathbf{v}^{(j)} = \sum_{j=1}^{n} \beta_j \lambda_j^2 \mathbf{v}^{(j)},$$

y, en general, $A^k\mathbf{x} = \sum_{j=1}^{n} \beta_j \lambda_j^k \mathbf{v}^{(j)}$.

Si λ_1^k se factoriza a partir de cada término en el lado derecho de la última ecuación, entonces

$$A^k\mathbf{x} = \lambda_1^k \sum_{j=1}^{n} \beta_j \left(\frac{\lambda_j}{\lambda_1}\right)^k \mathbf{v}^{(j)}.$$

Puesto que $|\lambda_1| > |\lambda_j|$, para todas $j = 2, 3, \ldots, n$, tenemos $\lim_{k\to\infty}(\lambda_j/\lambda_1)^k = 0$, y

$$\lim_{k\to\infty} A^k\mathbf{x} = \lim_{k\to\infty} \lambda_1^k \beta_1 \mathbf{v}^{(1)}. \tag{9.2}$$

La sucesión en la ecuación (9.2) converge a 0 si $|\lambda_1| < 1$ y diverge si $|\lambda_1| > 1$, siempre y cuando, por supuesto, $\beta_1 \neq 0$. Por consiguiente, las entradas en $A^k\mathbf{x}$ aumentarán con k si $|\lambda_1| > 1$ y tienden a 0 si $|\lambda_1| < 1$, tal vez, al resultar en desborde y subdesborde. Para cuidar esta posibilidad, escalamos las potencias de $A^k\mathbf{x}$ en una forma apropiada para garantizar que la cota en la ecuación (9.2) es finita y diferente de cero. El escalamiento comienza al seleccionar $\mathbf{x}$ como vector unitario $\mathbf{x}^{(0)}$ relativo a $\|\cdot\|_\infty$ y seleccionar un componente $x_{p_0}^{(0)}$ de $\mathbf{x}^{(0)}$ con

$$x_{p_0}^{(0)} = 1 = \|\mathbf{x}^{(0)}\|_\infty.$$

Sea $\mathbf{y}^{(1)} = A\mathbf{x}^{(0)}$ y defina $\mu^{(1)} = y_{p_0}^{(1)}$. Entonces

$$\mu^{(1)} = y_{p_0}^{(1)} = \frac{y_{p_0}^{(1)}}{x_{p_0}^{(0)}} = \frac{\beta_1\lambda_1 v_{p_0}^{(1)} + \sum_{j=2}^{n} \beta_j \lambda_j v_{p_0}^{(j)}}{\beta_1 v_{p_0}^{(1)} + \sum_{j=2}^{n} \beta_j v_{p_0}^{(j)}} = \lambda_1 \left[\frac{\beta_1 v_{p_0}^{(1)} + \sum_{j=2}^{n} \beta_j (\lambda_j/\lambda_1) v_{p_0}^{(j)}}{\beta_1 v_{p_0}^{(1)} + \sum_{j=2}^{n} \beta_j v_{p_0}^{(j)}}\right].$$

Sea p_1 el entero mínimo tal que

$$|y_{p_1}^{(1)}| = \|\mathbf{y}^{(1)}\|_\infty$$

y defina $\mathbf{x}^{(1)}$ mediante

$$\mathbf{x}^{(1)} = \frac{1}{y_{p_1}^{(1)}} \mathbf{y}^{(1)} = \frac{1}{y_{p_1}^{(1)}} A\mathbf{x}^{(0)}.$$

Entonces

$$x_{p_1}^{(1)} = 1 = \|\mathbf{x}^{(1)}\|_\infty.$$

Ahora defina

$$\mathbf{y}^{(2)} = A\mathbf{x}^{(1)} = \frac{1}{y_{p_1}^{(1)}} A^2\mathbf{x}^{(0)}$$

y

$$\mu^{(2)} = y_{p_1}^{(2)} = \frac{y_{p_1}^{(2)}}{x_{p_1}^{(1)}} = \frac{\left[\beta_1\lambda_1^2 v_{p_1}^{(1)} + \sum_{j=2}^n \beta_j\lambda_j^2 v_{p_1}^{(j)}\right] \Big/ y_{p_1}^{(1)}}{\left[\beta_1\lambda_1 v_{p_1}^{(1)} + \sum_{j=2}^n \beta_j\lambda_j v_{p_1}^{(j)}\right] \Big/ y_{p_1}^{(1)}}$$

$$= \lambda_1 \left[\frac{\beta_1 v_{p_1}^{(1)} + \sum_{j=2}^n \beta_j(\lambda_j/\lambda_1)^2 v_{p_1}^{(j)}}{\beta_1 v_{p_1}^{(1)} + \sum_{j=2}^n \beta_j(\lambda_j/\lambda_1) v_{p_1}^{(j)}}\right].$$

Sea p_2 el entero más pequeño con

$$|y_{p_2}^{(2)}| = \|\mathbf{y}^{(2)}\|_\infty$$

y defina

$$\mathbf{x}^{(2)} = \frac{1}{y_{p_2}^{(2)}}\mathbf{y}^{(2)} = \frac{1}{y_{p_2}^{(2)}} A\mathbf{x}^{(1)} = \frac{1}{y_{p_2}^{(2)} y_{p_1}^{(1)}} A^2\mathbf{x}^{(0)}.$$

De manera similar, defina las sucesiones de vectores $\{\mathbf{x}^{(m)}\}_{m=0}^\infty$ y $\{\mathbf{y}^{(m)}\}_{m=1}^\infty$ y una sucesión de escalares $\{\mu^{(m)}\}_{m=1}^\infty$ de manera inductiva mediante

$$\mathbf{y}^{(m)} = A\mathbf{x}^{(m-1)},$$

$$\mu^{(m)} = y_{p_{m-1}}^{(m)} = \lambda_1 \left[\frac{\beta_1 v_{p_{m-1}}^{(1)} + \sum_{j=2}^n (\lambda_j/\lambda_1)^m \beta_j v_{p_{m-1}}^{(j)}}{\beta_1 v_{p_{m-1}}^{(1)} + \sum_{j=2}^n (\lambda_j/\lambda_1)^{m-1} \beta_j v_{p_{m-1}}^{(j)}}\right] \tag{9.3}$$

y

$$\mathbf{x}^{(m)} = \frac{\mathbf{y}^{(m)}}{y_{p_m}^{(m)}} = \frac{A^m\mathbf{x}^{(0)}}{\prod_{k=1}^m y_{p_k}^{(k)}},$$

donde en cada paso, p_m se usa para representar el entero más pequeño para el que

$$|y_{p_m}^{(m)}| = \|\mathbf{y}^{(m)}\|_\infty.$$

Al examinar la ecuación (9.3), observamos que dado $|\lambda_j/\lambda_1| < 1$, para cada $j = 2, 3, \ldots, n$, $\lim_{m\to\infty} \mu^{(m)} = \lambda_1$, siempre y cuando $\mathbf{x}^{(0)}$ se seleccione de tal forma que $\beta_1 \neq 0$. Además, la sucesión de vectores $\{\mathbf{x}^{(m)}\}_{m=0}^\infty$ converge para un eigenvector asociado con λ_1 que tiene norma l_∞ igual a uno.

Ilustración La matriz

$$A = \begin{bmatrix} -2 & -3 \\ 6 & 7 \end{bmatrix}$$

tiene eigenvalores $\lambda_1 = 4$ y $\lambda_2 = 1$ con los eigenvectores correspondientes $\mathbf{v}_1 = (1, -2)^t$ y $\mathbf{v}_2 = (1, -1)^t$. Si comenzamos con el vector arbitrario $\mathbf{x}_0 = (1, 1)^t$ y multiplicamos por la matriz A, obtenemos

$$\mathbf{x}_1 = A\mathbf{x}_0 = \begin{bmatrix} -5 \\ 13 \end{bmatrix}, \qquad \mathbf{x}_2 = A\mathbf{x}_1 = \begin{bmatrix} -29 \\ 61 \end{bmatrix}, \qquad \mathbf{x}_3 = A\mathbf{x}_2 = \begin{bmatrix} -125 \\ 253 \end{bmatrix},$$

$$\mathbf{x}_4 = A\mathbf{x}_3 = \begin{bmatrix} -509 \\ 1021 \end{bmatrix}, \qquad \mathbf{x}_5 = A\mathbf{x}_4 = \begin{bmatrix} -2045 \\ 4093 \end{bmatrix}, \text{ y} \qquad \mathbf{x}_6 = A\mathbf{x}_5 = \begin{bmatrix} -8189 \\ 16381 \end{bmatrix}.$$

Como consecuencia, las aproximaciones para el eigenvalor dominante $\lambda_1 = 4$ son

$$\lambda_1^{(1)} = \frac{61}{13} = 4.6923, \qquad \lambda_1^{(2)} = \frac{253}{61} = 4.14754, \qquad \lambda_1^{(3)} = \frac{1021}{253} = 4.03557,$$

$$\lambda_1^{(4)} = \frac{4093}{1021} = 4.00881, \text{ y} \qquad \lambda_1^{(5)} = \frac{16381}{4093} = 4.00200.$$

Un eigenvector aproximado correspondiente a $\lambda_1^{(5)} = \dfrac{16381}{4093} = 4.00200$ es

$$\mathbf{x}_6 = \begin{bmatrix} -8189 \\ 16381 \end{bmatrix}, \text{ que, dividido entre } -8189, \text{ se normaliza a} \begin{bmatrix} 1 \\ -2.00037 \end{bmatrix} \approx \mathbf{v}_1.$$

■

El método de potencia tiene la desventaja de que es desconocido al principio si la matriz tiene un solo eigenvalor dominante. Tampoco se conoce cómo $\mathbf{x}^{(0)}$ debería seleccionarse para garantizar que su representación en términos de eigenvectores de la matriz contendrá una contribución diferente de cero de eigenvectores asociados con el eigenvalor dominante, si existiera.

El algoritmo 9.1 implementa el método de potencia.

ALGORITMO **9.1**

Método de potencia

Para aproximar el eigenvalor dominante y un eigenvector asociado de la matriz A $n \times n$ dado un vector $\mathbf{x}$ diferente a cero:

ENTRADA dimensión n; matriz A; vector $\mathbf{x}$; tolerancia TOL; número máximo de iteraciones N.

SALIDA aproximar el eigenvalor μ; aproximar el eigenvector $\mathbf{x}$ (con $||\mathbf{x}||_\infty = 1$) o un mensaje de que el número máximo de iteraciones fue excedido.

Paso 1 Determine $k = 1$.

Paso 2 Determine el entero más pequeño p con $1 \le p \le n$ y $|x_p| = ||\mathbf{x}||_\infty$.

Paso 3 Determine $\mathbf{x} = \mathbf{x}/x_p$.

Paso 4 Mientras ($k \le N$) haga los pasos 5–11.

  Paso 5 Determine $\mathbf{y} = A\mathbf{x}$.

  Paso 6 Determine $\mu = y_p$.

  Paso 7 Encuentre el entero más pequeño p con $1 \le p \le n$ y $|y_p| = \|\mathbf{y}\|_\infty$.

  Paso 8 Si $y_p = 0$ entonces SALIDA (‘Eigenvector’, $\mathbf{x}$);
  SALIDA (‘A tiene el eigenvalor 0, seleccione un nuevo vector $\mathbf{x}$ y reinicie’);
  PARE.

Paso 9 Determine $ERR = ||\mathbf{x} - (\mathbf{y}/y_p)||_\infty$;

$$\mathbf{x} = \mathbf{y}/y_p.$$

Paso 10 Si $ERR < TOL$ entonces SALIDA(μ, $\mathbf{x}$);

(*El procedimiento fue exitoso*.)
PARE.

Paso 11 Determine $k = k + 1$.

Paso 12 SALIDA ('El número máximo de iteraciones excedido');
(*El procedimiento no fue exitoso*.)
PARE. ■

Convergencia acelerada

Al seleccionar, en el paso 7, el entero más pequeño p_m para el que $|y_{p_m}^{(m)}| = \|\mathbf{y}^{(m)}\|_\infty$ garantizará, en general, que al final este índice se vuelve invariante. La velocidad a la que $\{\mu^{(m)}\}_{m=1}^\infty$ converge en λ_1 se determina mediante los radios $|\lambda_j/\lambda_1|^m$, para $j = 2, 3, \ldots, n$, y en particular por medio $|\lambda_2/\lambda_1|^m$. La velocidad de convergencia es $O(|\lambda_2/\lambda_1|^m)$ (consulte [IK], p. 148), por lo que existe una constante k, de tal forma que para m grande,

$$|\mu^{(m)} - \lambda_1| \approx k \left|\frac{\lambda_2}{\lambda_1}\right|^m,$$

lo cual implica que

$$\lim_{m\to\infty} \frac{|\mu^{(m+1)} - \lambda_1|}{|\mu^{(m)} - \lambda_1|} \approx \left|\frac{\lambda_2}{\lambda_1}\right| < 1.$$

La sucesión $\{\mu^{(m)}\}$ converge linealmente a λ_1, por lo que el procedimiento Δ^2 de Aitkens, que se analiza en la sección 2.5, se puede utilizar para acelerar la convergencia. Al implementar el procedimiento Δ^2 en el algoritmo 9.1 se logra modificar el algoritmo de acuerdo con lo siguiente:

Paso 1 Determine $k = 1$;
$\mu_0 = 0$;
$\mu_1 = 0$.

Paso 6 Determine $\mu = y_p$;

$$\hat{\mu} = \mu_0 - \frac{(\mu_1 - \mu_0)^2}{\mu - 2\mu_1 + \mu_0}.$$

Paso 10 Si $ERR < TOL$ y $k \geq 4$ entonces SALIDA ($\hat{\mu}$, $\mathbf{x}$);
PARE.

Paso 11 Determine $k = k + 1$;
$\mu_0 = \mu_1$;
$\mu_1 = \mu$.

En la actualidad, no es necesario que la matriz tenga diferentes eigenvalores para que el método de potencia converja. Si la matriz tiene un eigenvalor dominante único, λ_1, con multiplicidad r superior a 1 y $\mathbf{v}^{(1)}, \mathbf{v}^{(2)}, \ldots, \mathbf{v}^{(r)}$ son eigenvectores linealmente independientes asociados con λ_1, el procedimiento seguirá convergiendo en λ_1. En este caso, la sucesión de vectores $\{\mathbf{x}^{(m)}\}_{m=0}^\infty$, convergerá en un eigenvector de λ_1 en la norma l_∞ igual a uno que depende de la selección del vector inicial $\mathbf{x}^{(0)}$ y es una combinación lineal de $\mathbf{v}^{(1)}, \mathbf{v}^{(2)}, \ldots, \mathbf{v}^{(r)}$.

Ejemplo 1 Use el método de potencia para aproximar el eigenvalor dominante de la matriz

$$A = \begin{bmatrix} -4 & 14 & 0 \\ -5 & 13 & 0 \\ -1 & 0 & 2 \end{bmatrix}$$

y, a continuación, aplique el método Δ^2 de Aitkens para las aproximaciones para el eigenvalor de la matriz para acelerar la convergencia.

Solución Esta matriz tiene eigenvalores $\lambda_1 = 6, \lambda_2 = 3$, y $\lambda_3 = 2$, por lo que el método de potencia descrito en el algoritmo 9.1 convergerán. Si $\mathbf{x}^{(0)} = (1, 1, 1)^t$, entonces

$$\mathbf{y}^{(1)} = A\mathbf{x}^{(0)} = (10, 8, 1)^t,$$

por lo que

$$||\mathbf{y}^{(1)}||_\infty = 10, \quad \mu^{(1)} = y_1^{(1)} = 10, \quad \text{y} \quad \mathbf{x}^{(1)} = \frac{\mathbf{y}^{(1)}}{10} = (1, 0.8, 0.1)^t.$$

Al continuar de esta forma llegamos a los valores en la tabla 9.1, donde $\hat{\mu}^{(m)}$ representa la sucesión generada por el procedimiento Δ^2 de Aitkens. Una aproximación para el eigenvalor

Tabla 9.1

m	$(\mathbf{x}^{(m)})^t$	$\mu^{(m)}$	$\hat{\mu}^{(m)}$
0	$(1, 1, 1)$		
1	$(1, 0.8, 0.1)$	10	6.266667
2	$(1, 0.75, -0.111)$	7.2	6.062473
3	$(1, 0.730769, -0.188803)$	6.5	6.015054
4	$(1, 0.722200, -0.220850)$	6.230769	6.004202
5	$(1, 0.718182, -0.235915)$	6.111000	6.000855
6	$(1, 0.716216, -0.243095)$	6.054546	6.000240
7	$(1, 0.715247, -0.246588)$	6.027027	6.000058
8	$(1, 0.714765, -0.248306)$	6.013453	6.000017
9	$(1, 0.714525, -0.249157)$	6.006711	6.000003
10	$(1, 0.714405, -0.249579)$	6.003352	6.000000
11	$(1, 0.714346, -0.249790)$	6.001675	
12	$(1, 0.714316, -0.249895)$	6.000837	

dominante 6, en esta etapa es $\hat{\mu}^{(10)} = 6.000000$. El eigenvector unitario l_∞- aproximado para el eigenvalor 6 es $(\mathbf{x}^{(12)})^t = (1, 0.714316, -0.249895)^t$.

Aunque la aproximación para el eigenvalor es correcta para los lugares enumerados, la aproximación del eigenvector es considerablemente menos precisa para el eigenvector verdadero $(1, 5/7, -1/4)^t \approx (1, 0.714286, -0.25)^t$. ■

Matrices simétricas

Cuando A es simétrica, es posible hacer una variación en la selección de los vectores $\mathbf{x}^{(m)}$ y $\mathbf{y}^{(m)}$ y los escalares $\mu^{(m)}$ para mejorar significativamente el índice de convergencia de la sucesión $\{\mu^{(m)}\}_{m=1}^{\infty}$ para el eigenvalor dominante λ_1. De hecho, a pesar de que el índice de convergencia del método de potencia general es $O(|\lambda_2/\lambda_1|^m)$, el índice de convergencia

del procedimiento modificado que se dio en el algoritmo 9.2 para las matrices simétricas es $O(|\lambda_2/\lambda_1|^{2m})$. (Consulte [IK], p. 149 ff.) Puesto que la sucesión $\{\mu^{(m)}\}$ sigue siendo convergente, también puede aplicarse el procedimiento Δ^2 de Aitkens.

ALGORITMO
9.2

Método de potencia simétrica

Para aproximar el eigenvalor dominante y un eigenvector asociado de la matriz simétrica $n \times n$ A, dado un vector diferente de cero $\mathbf{x}$:

ENTRADA dimensión n; matriz A; vector $\mathbf{x}$; tolerancia TOL; número máximo de iteraciones N.

SALIDA aproxime el eigenvalor μ; aproxime el eigenvector $\mathbf{x}$ (con $\|\mathbf{x}\|_2 = 1$) o un mensaje de que el número máximo de iteraciones fue excedido.

Paso 1 Determine $k = 1$;
$\mathbf{x} = \mathbf{x}/\|\mathbf{x}\|_2$.

Paso 2 Mientras ($k \leq N$) haga los pasos 3–8.

Paso 3 Determine $\mathbf{y} = A\mathbf{x}$.

Paso 4 Determine $\mu = \mathbf{x}^t\mathbf{y}$.

Paso 5 Si $\|\mathbf{y}\|_2 = 0$, entonces SALIDA (‘Eigenvector’, $\mathbf{x}$);
SALIDA (‘A tiene un eigenvalor de 0, seleccione un nuevo vector $\mathbf{x}$ y reinicie’);
PARE.

Paso 6 Determine $ERR = \left\| \mathbf{x} - \dfrac{\mathbf{y}}{\|\mathbf{y}\|_2} \right\|_2$;
$\mathbf{x} = \mathbf{y}/\|\mathbf{y}\|_2$.

Paso 7 Si $ERR < TOL$ entonces SALIDA (μ, $\mathbf{x}$);
(*El procedimiento fue exitoso*.)
PARE.

Paso 8 Determine $k = k + 1$.

Paso 9 SALIDA (‘Número máximo de iteraciones excedido’);
(*El procedimiento no fue exitoso*.)
PARE. ■

Ejemplo 2 Aplique tanto el método de potencia como el de potencia simétrica a la matriz

$$A = \begin{bmatrix} 4 & -1 & 1 \\ -1 & 3 & -2 \\ 1 & -2 & 3 \end{bmatrix},$$

usando el método Δ^2 de Aitkens para acelerar la convergencia.

Solución Esta matriz tiene eigenvalores $\lambda_1 = 6$, $\lambda_2 = 3$ y $\lambda_3 = 1$. Un eigenvector para el eigenvalor 6 es $(1, -1, 1)^t$. La aplicación del método de potencia a esta matriz con vector inicial $(1, 0, 0)^t$ da los valores de la tabla 9.2.

Tabla 9.2

m	$(\mathbf{y}^{(m)})^t$	$\mu^{(m)}$	$\hat{\mu}^{(m)}$	$(\mathbf{x}^{(m)})^t$ con $\|\mathbf{x}^{(m)}\|_\infty = 1$
0				$(1, 0, 0)$
1	$(4, -1, 1)$	4		$(1, -0.25, 0.25)$
2	$(4.5, -2.25, 2.25)$	4.5	7	$(1, -0.5, 0.5)$
3	$(5, -3.5, 3.5)$	5	6.2	$(1, -0.7, 0.7)$
4	$(5.4, -4.5, 4.5)$	5.4	6.047617	$(1, -0.833\bar{3}, 0.833\bar{3})$
5	$(5.66\bar{6}, -5.166\bar{6}, 5.166\bar{6})$	$5.66\bar{6}$	6.011767	$(1, -0.911765, 0.911765)$
6	$(5.823529, -5.558824, 5.558824)$	5.823529	6.002931	$(1, -0.954545, 0.954545)$
7	$(5.909091, -5.772727, 5.772727)$	5.909091	6.000733	$(1, -0.976923, 0.976923)$
8	$(5.953846, -5.884615, 5.884615)$	5.953846	6.000184	$(1, -0.988372, 0.988372)$
9	$(5.976744, -5.941861, 5.941861)$	5.976744		$(1, -0.994163, 0.994163)$
10	$(5.988327, -5.970817, 5.970817)$	5.988327		$(1, -0.997076, 0.997076)$

Ahora aplicaremos el método de potencia simétrica a esta matriz con el mismo vector inicial $(1, 0, 0)^t$. Los primeros pasos son

$$\mathbf{x}^{(0)} = (1, 0, 0)^t, \quad A\mathbf{x}^{(0)} = (4, -1, 1)^t,\ M^{(1)} = 4,$$

y

$$\mathbf{x}^{(1)} = \frac{1}{||A\mathbf{x}^{(0)}||_2} \cdot A\mathbf{x}^{(0)} = (0.942809, -0.235702, 0.235702)^t.$$

Las entradas restantes se muestran en la tabla 9.3.

Tabla 9.3

m	$(\mathbf{y}^{(m)})^t$	$\mu^{(m)}$	$\hat{\mu}^{(m)}$	$(\mathbf{x}^{(m)})^t$ con $\|\mathbf{x}^{(m)}\|_2 = 1$
0	$(1, 0, 0)$			$(1, 0, 0)$
1	$(4, -1, 1)$	4	7	$(0.942809, -0.235702, 0.235702)$
2	$(4.242641, -2.121320, 2.121320$	5	6.047619	$(0.816497, -0.408248, 0.408248)$
3	$(4.082483, -2.857738, 2.857738)$	5.666667	6.002932	$(0.710669, -0.497468, 0.497468)$
4	$(3.837613, -3.198011, 3.198011)$	5.909091	6.000183	$(0.646997, -0.539164, 0.539164)$
5	$(3.666314, -3.342816, 3.342816)$	5.976744	6.000012	$(0.612836, -0.558763, 0.558763)$
6	$(3.568871, -3.406650, 3.406650)$	5.994152	6.000000	$(0.595247, -0.568190, 0.568190)$
7	$(3.517370, -3.436200, 3.436200)$	5.998536	6.000000	$(0.586336, -0.572805, 0.572805)$
8	$(3.490952, -3.450359, 3.450359)$	5.999634		$(0.581852, -0.575086, 0.575086)$
9	$(3.477580, -3.457283, 3.457283)$	5.999908		$(0.579603, -0.576220, 0.576220)$
10	$(3.470854, -3.460706, 3.460706)$	5.999977		$(0.578477, -0.576786, 0.576786)$

El método de potencia simétrica da una convergencia considerablemente más rápida para esta matriz que el método de potencia. Las aproximaciones del eigenvector en el método de potencia converge en $(1, -1, 1)^t$, un vector con norma unitaria en l_∞. En el método de potencia simétrica, la convergencia es el vector paralelo $(\sqrt{3}/3, -\sqrt{3}/3, \sqrt{3}/3)^t$, que tiene la norma unitaria en l_2. ▪

Si λ es un número real que aproxima un eigenvalor de una matriz simétrica A y $\mathbf{x}$ es un eigenvector asociado aproximado, entonces $A\mathbf{x} - \lambda\mathbf{x}$ es aproximadamente el vector cero. El siguiente teorema relaciona la norma de este vector para la precisión del eigenvalor λ.

Teorema 9.19 Suponga que A es una matriz simétrica $n \times n$ con eigenvalores $\lambda_1, \lambda_2, \ldots, \lambda_n$. Si $\|A\mathbf{x} - \lambda\mathbf{x}\|_2 < \varepsilon$ para algunos números reales λ y vector $\mathbf{x}$ con $\|\mathbf{x}\|_2 = 1$. Entonces

$$\min_{1 \le j \le n} |\lambda_j - \lambda| < \varepsilon.$$

Demostración Suponga que $\mathbf{v}^{(1)}, \mathbf{v}^{(2)}, \ldots, \mathbf{v}^{(n)}$ forman un conjunto ortonormal de eigenvectores asociados de A, respectivamente, con los eigenvalores $\lambda_1, \lambda_2, \ldots, \lambda_n$. Mediante los teoremas 9.5 y 9.3, $\mathbf{x}$ se puede expresar, para algún conjunto único de constantes $\beta_1, \beta_2, \ldots, \beta_n$, como

$$\mathbf{x} = \sum_{j=1}^{n} \beta_j \mathbf{v}^{(j)}.$$

Por lo tanto,

$$\|A\mathbf{x} - \lambda\mathbf{x}\|_2^2 = \left\| \sum_{j=1}^{n} \beta_j(\lambda_j - \lambda)\mathbf{v}^{(j)} \right\|_2^2 = \sum_{j=1}^{n} |\beta_j|^2 |\lambda_j - \lambda|^2 \geq \min_{1\leq j\leq n} |\lambda_j - \lambda|^2 \sum_{j=1}^{n} |\beta_j|^2.$$

Pero

$$\sum_{j=1}^{n} |\beta_j|^2 = \|\mathbf{x}\|_2^2 = 1, \quad \text{por lo que} \quad \varepsilon \geq \|A\mathbf{x} - \lambda\mathbf{x}\|_2 > \min_{1\leq j\leq n} |\lambda_j - \lambda|.$$ ■

Método de potencia inversa

El **método de potencia inversa** es una modificación del método de potencia que da una convergencia más rápida. Se usa para determinar el eigenvalor de A que está más cerca de un número específico q.

Suponga que la matriz A tiene eigenvalores $\lambda_1, \ldots, \lambda_n$ con eigenvectores linealmente independientes $\mathbf{v}^{(1)}, \ldots, \mathbf{v}^{(n)}$. Los eigenvalores de $(A - qI)^{-1}$, donde $q \neq \lambda_i$, para $i = 1, 2, \ldots, n$, son

$$\frac{1}{\lambda_1 - q}, \quad \frac{1}{\lambda_2 - q}, \ldots, \frac{1}{\lambda_n - q},$$

con estos mismos eigenvectores $\mathbf{v}^{(1)}, \mathbf{v}^{(2)}, \ldots, \mathbf{v}^{(n)}$. (Consulte el ejercicio 17 de la sección 7.2.)

Al aplicar el método de potencia a $(A - qI)^{-1}$ da

$$\mathbf{y}^{(m)} = (A - qI)^{-1}\mathbf{x}^{(m-1)},$$

$$\mu^{(m)} = y_{p_{m-1}}^{(m)} = \frac{y_{p_{m-1}}^{(m)}}{x_{p_{m-1}}^{(m-1)}} = \frac{\sum_{j=1}^{n} \beta_j \dfrac{1}{(\lambda_j - q)^m} v_{p_{m-1}}^{(j)}}{\sum_{j=1}^{n} \beta_j \dfrac{1}{(\lambda_j - q)^{m-1}} v_{p_{m-1}}^{(j)}}, \tag{9.4}$$

y

$$\mathbf{x}^{(m)} = \frac{\mathbf{y}^{(m)}}{y_{p_m}^{(m)}},$$

donde, en cada paso, p_m representa el entero más pequeño para el que $|y_{p_m}^{(m)}| = ||\mathbf{y}^{(m)}||_\infty$. La sucesión $\{\mu^{(m)}\}$ en la ecuación (9.4) converge en $1/(\lambda_k - q)$, donde

$$\frac{1}{|\lambda_k - q|} = \max_{1\leq i\leq n} \frac{1}{|\lambda_i - q|}$$

y $\lambda_k \approx q + 1/\mu^{(m)}$ es el eigenvalor de A más cercano a q.

Conociendo k, la ecuación (9.4) se puede escribir como

$$\mu^{(m)} = \frac{1}{\lambda_k - q}\left[\frac{\beta_k v_{p_{m-1}}^{(k)} + \sum_{\substack{j=1 \\ j\neq k}}^{n} \beta_j \left[\frac{\lambda_k - q}{\lambda_j - q}\right]^m v_{p_{m-1}}^{(j)}}{\beta_k v_{p_{m-1}}^{(k)} + \sum_{\substack{j=1 \\ j\neq k}}^{n} \beta_j \left[\frac{\lambda_k - q}{\lambda_j - q}\right]^{m-1} v_{p_{m-1}}^{(j)}}\right]. \tag{9.5}$$

Por lo tanto, la selección de q determina la convergencia, siempre y cuando $1/(\lambda_k - q)$ sea un único eigenvalor dominante de $(A - qI)^{-1}$ (a pesar de que puede ser un eigenvalor múltiple). Mientras q está más cerca de un eigenvalor λ_k, más rápida será la convergencia ya que la convergencia es de orden

$$O\left(\left|\frac{(\lambda - q)^{-1}}{(\lambda_k - q)^{-1}}\right|^m\right) = O\left(\left|\frac{(\lambda_k - q)}{(\lambda - q)}\right|^m\right),$$

donde λ representa el eigenvalor de A que es el segundo más cercano a q.

El vector $y^{(m)}$ se obtiene al resolver el sistema lineal

$$(A - qI)\mathbf{y}^{(m)} = \mathbf{x}^{(m-1)}.$$

En general, se usa la eliminación gaussiana con pivoteo pero, como en el caso de la factorización LU, los multiplicadores se pueden guardar para reducir el cálculo. La selección de q puede basarse en el teorema del círculo de Geršgorin o en otros medios de localización de un eigenvalor.

El algoritmo 9.3 calcula q a partir de una aproximación inicial para el eigenvalor $\mathbf{x}^{(0)}$ mediante

$$q = \frac{\mathbf{x}^{(0)t} A\mathbf{x}^{(0)}}{\mathbf{x}^{(0)t}\mathbf{x}^{(0)}}.$$

Esta selección de q resulta de la observación de que si $\mathbf{x}$ es un eigenvector de A respecto al eigenvalor λ, entonces $A\mathbf{x} = \lambda\mathbf{x}$. Por lo que, $\mathbf{x}^t A\mathbf{x} = \lambda\mathbf{x}^t\mathbf{x}$ y

$$\lambda = \frac{\mathbf{x}^t A\mathbf{x}}{\mathbf{x}^t\mathbf{x}} = \frac{\mathbf{x}^t A\mathbf{x}}{\|\mathbf{x}\|_2^2}.$$

Si q está cerca de un eigenvalor, la convergencia sería bastante rápida, pero se debería usar una técnica de pivoteo en el paso 6 para evitar la contaminación por error de redondeo.

A menudo, se usa el algoritmo 9.3 para aproximar un eigenvalor cuando se conoce un eigenvalor q aproximado.

ALGORITMO 9.3

Método de potencia inversa

Para aproximar un eigenvalor y un eigenvector asociado de la matriz A $n \times n$, dado un vector $\mathbf{x}$ diferente de cero:

ENTRADA dimensión n; matriz A; vector $\mathbf{x}$; tolerancia TOL; número máximo de iteraciones N.

SALIDA aproxima el eigenvalor μ; aproxima el eigenvector $\mathbf{x}$ (con $\|\mathbf{x}\|_\infty = 1$) o un mensaje de que el número máximo de iteraciones fue excedido.

Paso 1 Determine $q = \dfrac{\mathbf{x}^t A\mathbf{x}}{\mathbf{x}^t\mathbf{x}}$.

Paso 2 Determine $k = 1$.

Paso 3 Encuentre el entero más pequeño p con $1 \leq p \leq n$ y $|x_p| = \|\mathbf{x}\|_\infty$.

Paso 4 Determine $\mathbf{x} = \mathbf{x}/x_p$.

Paso 5 Mientras $(k \leq N)$ haga los pasos 6–12.

Paso 6 Resuelva el sistema lineal $(A - qI)\mathbf{y} = \mathbf{x}$.

Paso 7 Si el sistema no tiene una solución única, entonces
SALIDA ('q es un eigenvalor', q);
PARE.

Paso 8 Determine $\mu = y_p$.

Paso 9 Encuentre el entero más pequeño p con $1 \leq p \leq n$ y $|y_p| = \|\mathbf{y}\|_\infty$.

Paso 10 Determine $ERR = \left\|\mathbf{x} - (\mathbf{y}/y_p)\right\|_\infty$;

$$\mathbf{x} = \mathbf{y}/y_p.$$

Paso 11 Si $ERR < TOL$ entonces determine $\mu = (1/\mu) + q$;
SALIDA $(\mu, \mathbf{x})$;
(*El procedimiento fue exitoso*.)
PARE.

Paso 12 Determine $k = k + 1$.

Paso 13 SALIDA ('Número máximo de iteraciones excedido');
(*El procedimiento no fue exitoso*.)
PARE. ■

La convergencia del método de potencia inversa es lineal, por lo que, de nuevo, el método Δ^2 de Aitkens puede usarse para acelerar la convergencia. El siguiente ejemplo ilustra la rápida convergencia del método de potencia inversa si q está cerca de un eigenvalor.

Ejemplo 3 Aplique el método de potencia inversa con $\mathbf{x}^{(0)} = (1, 1, 1)^t$ a la matriz

$$A = \begin{bmatrix} -4 & 14 & 0 \\ -5 & 13 & 0 \\ -1 & 0 & 2 \end{bmatrix} \quad \text{con} \quad q = \frac{\mathbf{x}^{(0)t}A\mathbf{x}^{(0)}}{\mathbf{x}^{(0)t}\mathbf{x}^{(0)}} = \frac{19}{3}$$

y use el método Δ^2 de Aitkens para acelerar la convergencia.

Solución El método de potencia se aplicó a esta matriz en el ejemplo 1 usando el vector inicial $\mathbf{x}^{(0)} = (1, 1, 1)^t$. Éste nos dio el eigenvalor $\mu^{(12)} = 6.000837$ y el eigenvector $(\mathbf{x}^{(12)})^t = (1, 0.714316, -0.249895)^t$.

Para el método de potencia inversa, consideramos

$$A - qI = \begin{bmatrix} -\frac{31}{3} & 14 & 0 \\ -5 & \frac{20}{3} & 0 \\ -1 & 0 & -\frac{13}{3} \end{bmatrix}.$$

Con $\mathbf{x}^{(0)} = (1, 1, 1)^t$, el método encuentra primero $\mathbf{y}^{(1)}$ al resolver $(A - qI)\mathbf{y}^{(1)} = \mathbf{x}^{(0)}$. Esto da

$$\mathbf{y}^{(1)} = \left(-\frac{33}{5}, -\frac{24}{5}, \frac{84}{65}\right)^t = (-6.6, -4.8, 1.292\overline{307692})^t.$$

Por lo que

$$||\mathbf{y}^{(1)}||_\infty = 6.6, \quad \mathbf{x}^{(1)} = \frac{1}{-6.6}\mathbf{y}^{(1)} = (1, 0.7272727, -0.1958042)^t,$$

y

$$\mu^{(1)} = -\frac{1}{6.6} + \frac{19}{3} = 6.1818182.$$

Los resultados subsiguientes se incluyen en la tabla 9.4 y la columna derecha lista los resultados del método Δ^2 de Aitkens aplicado a $\mu^{(m)}$. Estos son resultados claramente superiores a los obtenidos con el método de potencia. ■

Tabla 9.4

m	$\mathbf{x}^{(m)t}$	$\mu^{(m)}$	$\hat{\mu}^{(m)}$
0	$(1, 1, 1)$		
1	$(1, 0.7272727, -0.1958042)$	6.1818182	6.000098
2	$(1, 0.7155172, -0.2450520)$	6.0172414	6.000001
3	$(1, 0.7144082, -0.2495224)$	6.0017153	6.000000
4	$(1, 0.7142980, -0.2499534)$	6.0001714	6.000000
5	$(1, 0.7142869, -0.2499954)$	6.0000171	
6	$(1, 0.7142858, -0.2499996)$	6.0000017	

Si A es simétrica, entonces, para cualquier número real q, la matriz $(A-qI)^{-1}$ también es simétrica; por lo que el método de potencia simétrica, algoritmo 9.2, se puede aplicar a $(A-qI)^{-1}$ para acelerar la convergencia en

$$O\left(\left|\frac{\lambda_k - q}{\lambda - q}\right|^{2m}\right).$$

Métodos de deflación

Existen numerosas técnicas para obtener aproximaciones para los otros eigenvalores de una matriz, una vez que se ha calculado una aproximación al eigenvalor dominante. Restringiremos nuestra presentación a las **técnicas de deflación**.

Las técnicas de deflación implican la formación de una nueva matriz B, cuyos eigenvalores sean iguales a los de A, excepto que el eigenvalor dominante de A es reemplazado en B por el eigenvalor 0. El siguiente resultado justifica el procedimiento. La demostración de este teorema se puede encontrar en [Wil2], p. 596.

Teorema 9.20 Suponga que $\lambda_1, \lambda_2, \ldots, \lambda_n$ son eigenvalores de A con eigenvectores asociados $\mathbf{v}^{(1)}, \mathbf{v}^{(2)}, \ldots \mathbf{v}^{(n)}$ y λ_1 tiene multiplicidad 1. Sea $\mathbf{x}$ un vector con $\mathbf{x}^t\mathbf{v}^{(1)} = 1$. Entonces la matriz

$$B = A - \lambda_1 \mathbf{v}^{(1)}\mathbf{x}^t$$

tiene eigenvalores $0, \lambda_2, \lambda_3, \ldots, \lambda_n$ con eigenvectores asociados $\mathbf{v}^{(1)}, \mathbf{w}^{(2)}, \mathbf{w}^{(3)}, \ldots, \mathbf{w}^{(n)}$, donde $\mathbf{v}^{(i)}$ y $\mathbf{w}^{(i)}$ están relacionados por medio de la ecuación

$$\mathbf{v}^{(i)} = (\lambda_i - \lambda_1)\mathbf{w}^{(i)} + \lambda_1(\mathbf{x}^t\mathbf{w}^{(i)})\mathbf{v}^{(1)}, \tag{9.6}$$

para cada $i = 2, 3, \ldots, n$. ■

Existen muchas selecciones para el vector $\mathbf{x}$ que podrían usarse en el teorema 9.20. La **deflación de Wielandt** inicia con la definición

$$\mathbf{x} = \frac{1}{\lambda_1 v_i^{(1)}}(a_{i1}, a_{i2}, \ldots, a_{in})^t, \tag{9.7}$$

Helmut Wielandt (1910–2001) trabajó originalmente en grupos de permutación, pero durante la Segunda Guerra Mundial se comprometió con la investigación en meteorología, criptología y aerodinámica. Esto implicaba problemas de vibración que requerían el cálculo de eigenvalores asociados con ecuaciones y matrices diferenciales.

donde $v_i^{(1)}$ es una coordenada diferente de cero del eigenvector $\mathbf{v}^{(1)}$ y los valores $a_{i1}, a_{i2}, \ldots, a_{in}$ son las entradas en la i-ésima fila de A.

Con esta definición,

$$\mathbf{x}^t\mathbf{v}^{(1)} = \frac{1}{\lambda_1 v_i^{(1)}}[a_{i1}, a_{i2}, \ldots, a_{in}](v_1^{(1)}, v_2^{(1)}, \ldots, v_n^{(1)})^t = \frac{1}{\lambda_1 v_i^{(1)}}\sum_{j=1}^{n} a_{ij}v_j^{(1)},$$

donde la suma es la i-ésima coordenada del producto $A\mathbf{v}^{(1)}$. Puesto que $A\mathbf{v}^{(1)} = \lambda_1\mathbf{v}^{(1)}$, tenemos

$$\sum_{j=1}^{n} a_{ij}v_j^{(1)} = \lambda_1 v_i^{(1)},$$

lo cual implica que

$$\mathbf{x}^t\mathbf{v}^{(1)} = \frac{1}{\lambda_1 v_i^{(1)}}(\lambda_1 v_i^{(1)}) = 1.$$

Por lo tanto, $\mathbf{x}$ satisface la hipótesis del teorema 9.20. Además (consulte el ejercicio 25), la i-ésima fila de $B = A - \lambda_1\mathbf{v}^{(1)}\mathbf{x}^t$ contiene completamente entradas cero.

Si $\lambda \neq 0$ es un eigenvalor con un eigenvector asociado $\mathbf{w}$, la relación $B\mathbf{w} = \lambda\mathbf{w}$ implica que la i-ésima coordenada de $\mathbf{w}$ también debe ser cero. Por consiguiente, la i-ésima columna de la matriz B no contribuye al producto $B\mathbf{w} = \lambda\mathbf{w}$. Por lo tanto, la matriz B se puede reemplazar con una matriz B' $(n-1)\times(n-1)$ obtenida al eliminar la i-ésima fila y la i-ésima columna de B. La matriz B' tiene eigenvalores $\lambda_2, \lambda_3, \ldots, \lambda_n$.

Si $|\lambda_2| > |\lambda_3|$, el método de potencia se aplica nuevamente a la matriz B' para determinar este eigenvalor dominante y un eigenvector $\mathbf{w}^{(2)'}$, asociado con λ_2, respecto a la matriz B'. Para encontrar el eigenvector asociado $\mathbf{w}^{(2)}$ para la matriz B, inserte una coordinada cero entre las coordenadas $w_{i-1}^{(2)'}$ y $w_i^{(2)'}$ del vector $(n-1)$ dimensional $\mathbf{w}^{(2)'}$ y, después, calcule $\mathbf{v}^{(2)}$ con la ecuación (9.6).

Ejemplo 4 La matriz

$$A = \begin{bmatrix} 4 & -1 & 1 \\ -1 & 3 & -2 \\ 1 & -2 & 3 \end{bmatrix}$$

tiene el eigenvalor dominante $\lambda_1 = 6$ con eigenvector unitario asociado $\mathbf{v}^{(1)} = (1, -1, 1)^t$. Suponga que conocemos este eigenvalor dominante y aplique la deflación para aproximar los otros eigenvalores y eigenvectores.

Solución El procedimiento para obtener un segundo eigenvalor λ_2 procede de acuerdo con lo siguiente:

$$\mathbf{x} = \frac{1}{6}\begin{bmatrix} 4 \\ -1 \\ 1 \end{bmatrix} = \left(\frac{2}{3}, -\frac{1}{6}, \frac{1}{6}\right)^t,$$

$$\mathbf{v}^{(1)}\mathbf{x}^t = \begin{bmatrix} 1 \\ -1 \\ 1 \end{bmatrix}\left[\ \tfrac{2}{3}, \quad -\tfrac{1}{6}, \quad \tfrac{1}{6}\ \right] = \begin{bmatrix} \frac{2}{3} & -\frac{1}{6} & \frac{1}{6} \\ -\frac{2}{3} & \frac{1}{6} & -\frac{1}{6} \\ \frac{2}{3} & -\frac{1}{6} & \frac{1}{6} \end{bmatrix},$$

y

$$B = A - \lambda_1 \mathbf{v}^{(1)}\mathbf{x}^t = \begin{bmatrix} 4 & -1 & 1 \\ -1 & 3 & -2 \\ 1 & -2 & 3 \end{bmatrix} - 6 \begin{bmatrix} \frac{2}{3} & -\frac{1}{6} & \frac{1}{6} \\ -\frac{2}{3} & \frac{1}{6} & -\frac{1}{6} \\ \frac{2}{3} & -\frac{1}{6} & \frac{1}{6} \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 \\ 3 & 2 & -1 \\ -3 & -1 & 2 \end{bmatrix}.$$

Al eliminar la primera fila y la primera columna obtenemos

$$B' = \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix},$$

que tiene eigenvalores $\lambda_2 = 3$ y $\lambda_3 = 1$. Para $\lambda_2 = 3$, el eigenvector $\mathbf{w}^{(2)'}$ se puede obtener al resolver el sistema lineal

$$(B' - 3I)\mathbf{w}^{(2)'} = \mathbf{0}, \quad \text{que resulta en} \quad \mathbf{w}^{(2)'} = (1, -1)^t.$$

Al agregar un cero al primer componente obtenemos $\mathbf{w}^{(2)} = (0, 1, -1)^t$, y, a partir de la ecuación (9.6), tenemos el eigenvector $\mathbf{v}^{(2)}$ de A correspondiente a $x_2 = 3$:

$$\begin{aligned} \mathbf{v}^{(2)} &= (\lambda_2 - \lambda_1)\mathbf{w}^{(2)} + \lambda_1(\mathbf{x}^t\mathbf{w}^{(2)})\mathbf{v}^{(1)} \\ &= (3-6)(0, 1, -1)^t + 6\left[\left(\frac{2}{3}, -\frac{1}{6}, \frac{1}{6}\right)(0, 1, -1)^t\right](1, -1, 1)^t = (-2, -1, 1)^t. \end{aligned}$$ ■

Aunque este proceso de deflación se puede usar para encontrar las aproximaciones para todos los eigenvalores y eigenvectores de una matriz, el proceso es susceptible al error de redondeo. Después de usar la deflación para aproximar el eigenvalor de una matriz, la aproximación debería utilizarse como valor inicial para el método de potencia inversa aplicado a la matriz original. Esto garantizará la convergencia para un eigenvalor de la matriz original, no la de las matrices reducidas, que probablemente contiene errores. Cuando se requieren todos los eigenvalores de una matriz, deberían usarse las técnicas consideradas en la sección 9.5, con base en transformaciones de similitud.

Cerramos esta sección con el algoritmo 9.4, que calcula el segundo eigenvalor dominante y el eigenvector asociado para una matriz, una vez que se ha determinado el eigenvalor dominante y el eigenvector asociado.

ALGORITMO **9.4**

Deflación de Wielandt

Para aproximar el segundo eigenvalor más dominante y un eigenvector asociado de la matriz A $n \times n$ dada una aproximación λ para el eigenvalor dominante, una aproximación $\mathbf{v}$ para un eigenvector asociado y un vector $\mathbf{x} \in \mathbb{R}^{n-1}$:

ENTRADA dimensión n; matriz A; eigenvalor λ aproximado con eigenvector $\mathbf{v} \in \mathbb{R}^n$; vector $\mathbf{x} \in \mathbb{R}^{n-1}$; tolerancia TOL; número máximo de iteraciones N.

SALIDA eigenvalor μ aproximado; eigenvector aproximado $\mathbf{u}$ o un mensaje de que el método falla.

Paso 1 Sea i el entero más pequeño con $1 \le i \le n$ y $|v_i| = \max_{1 \le j \le n} |v_j|$.

Paso 2 Si $i \ne 1$ entonces
    para $k = 1, \ldots, i-1$
        para $j = 1, \ldots, i-1$
            determine $b_{kj} = a_{kj} - \dfrac{v_k}{v_i} a_{ij}$.

Paso 3 Si $i \neq 1$ y $i \neq n$ entonces
para $k = i, \ldots, n-1$
para $j = 1, \ldots, i-1$

$$\text{determine } b_{kj} = a_{k+1,j} - \frac{v_{k+1}}{v_i} a_{ij};$$

$$b_{jk} = a_{j,k+1} - \frac{v_j}{v_i} a_{i,k+1}.$$

Paso 4 Si $i \neq n$ entonces
para $k = i, \ldots, n-1$
para $j = i, \ldots, n-1$

$$\text{determine } b_{kj} = a_{k+1,j+1} - \frac{v_{k+1}}{v_i} a_{i,j+1}.$$

Paso 5 Realice el método de potencia en la matriz $(n-1) \times (n-1)$ $B' = (b_{kj})$ con $\mathbf{x}$ como aproximación inicial.

Paso 6 Si el método falla, entonces SALIDA ('El método falla');
PARE.
si μ es el eigenvalor aproximado
$\mathbf{w}' = (w_1', \ldots, w_{n-1}')^t$ el eigenvector aproximado.

Paso 7 Si $i \neq 1$ entonces para $k = 1, \ldots, i-1$ determine $w_k = w_k'$.

Paso 8 Determine $w_i = 0$.

Paso 9 Si $i \neq n$ entonces para $k = i+1, \ldots, n$ determine $w_k = w_{k-1}'$.

Paso 10 Para $k = 1, \ldots, n$

$$\text{determine } u_k = (\mu - \lambda) w_k + \left(\sum_{j=1}^{n} a_{ij} w_j \right) \frac{v_k}{v_i}.$$

(*Calcule el eigenvector con la ecuación (9.6).*)

Paso 11 SALIDA $(\mu, \mathbf{u})$; (*El procedimiento fue exitoso* .)
PARE. ■

La sección Conjunto de ejercicios 9.3 está disponible en línea. Encuentre la ruta de acceso en las páginas preliminares.

9.4 Método de Householder

En la sección 9.5 usaremos el método QR para reducir una matriz tridiagonal simétrica en una matriz similar que es casi diagonal. Las entradas diagonales de la matriz reducida son aproximaciones para los eigenvalores de la matriz dada. En esta sección presentamos un método concebido por Alston Householder para reducir una matriz simétrica arbitraria en una matriz tridiagonal similar. A pesar de que existe una conexión entre los problemas que estamos resolviendo en estas dos secciones, el método de Householder tiene una aplicación tan amplia en áreas diferentes a la aproximación de eigenvalores que merece un trato especial.

Alston Householder (1904–1993) realizó investigaciones en biología matemática antes de ser director del Laboratorio Nacional de Oak Ridge en Tennessee en 1948. Comenzó a trabajar en la solución de sistemas lineales en la década de 1950, cuando se desarrollaron estos métodos.

El método de Householder se usa para encontrar una matriz tridiagonal simétrica B que es similar a una matriz simétrica A determinada. El teorema 9.16 implica que A es similar a la matriz diagonal D, ya que existe una matriz ortogonal Q con la propiedad de que $D = Q^{-1}AQ = Q^tAQ$. Puesto que, en general, la matriz Q (y por consiguiente, D) es difícil de calcular, el método de Householder ofrece un compromiso. Después de haber implementado el método de Householder, es posible usar métodos eficientes, como el algoritmo QR, para aproximar con exactitud los eigenvalores de la matriz tridiagonal simétrica resultante.

Transformaciones de Householder

Definición 9.21 Sea $\mathbf{w} \in \mathbb{R}^n$ con $\mathbf{w}^t\mathbf{w} = 1$. Entonces la matriz $n \times n$

$$P = I - 2\mathbf{w}\mathbf{w}^t$$

recibe el nombre de **transformación de Householder**. ■

Las transformaciones de Householder se usan para los bloques externos de entradas cero en vectores o columnas de matrices de manera en extremo estable respecto al error de redondeo. (Consulte [Wil2], pp. 152–162, para mayor análisis.) Las propiedades de las transformaciones se dan en el siguiente teorema.

Teorema 9.22 Si una transformación de Householder, $P = I - 2\mathbf{w}\mathbf{w}^t$, es simétrica y ortogonal, entonces $P^{-1} = P$.

Demostración

$$(\mathbf{w}\mathbf{w}^t)^t = (\mathbf{w}^t)^t\mathbf{w}^t = \mathbf{w}\mathbf{w}^t$$

y de

$$P^t = (I - 2\mathbf{w}\mathbf{w}^t)^t = I - 2\mathbf{w}\mathbf{w}^t = P.$$

Además, $\mathbf{w}^t\mathbf{w} = 1$, por lo que

$$\begin{aligned} PP^t = (I - 2\mathbf{w}\mathbf{w}^t)(I - 2\mathbf{w}\mathbf{w}^t) &= I - 2\mathbf{w}\mathbf{w}^t - 2\mathbf{w}\mathbf{w}^t + 4\mathbf{w}\mathbf{w}^t\mathbf{w}\mathbf{w}^t \\ &= I - 4\mathbf{w}\mathbf{w}^t + 4\mathbf{w}\mathbf{w}^t = I, \end{aligned}$$

y $P^{-1} = P^t = P$. ■

El método de Householder comienza determinando una transformación $P^{(1)}$ tal que $A^{(2)} = P^{(1)}AP^{(1)}$ tiene entradas ceros fuera de la primera columna de A, comenzando con la tercera fila, es decir,

$$a_{j1}^{(2)} = 0, \quad \text{para cada } j = 3, 4, \ldots, n. \tag{9.8}$$

Por simetría, también tenemos $a_{1j}^{(2)} = 0$.

Ahora seleccionamos un vector $\mathbf{w} = (w_1, w_2, \ldots, w_n)^t$ de tal forma que $\mathbf{w}^t\mathbf{w} = 1$, la ecuación (9.8) se mantiene y en la matriz

$$A^{(2)} = P^{(1)}AP^{(1)} = (I - 2\mathbf{w}\mathbf{w}^t)A(I - 2\mathbf{w}\mathbf{w}^t),$$

tenemos $a_{11}^{(2)} = a_{11}$ y $a_{j1}^{(2)} = 0$, para cada $j = 3, 4, \ldots, n$. Esta selección impone n condiciones en los n valores desconocidos $w_1, w_2, \ldots, w_n$.

Al establecer $w_1 = 0$ garantizamos que $a_{11}^{(2)} = a_{11}$. Queremos

$$P^{(1)} = I - 2\mathbf{w}\mathbf{w}^t$$

para satisfacer

$$P^{(1)}(a_{11}, a_{21}, a_{31}, \ldots, a_{n1})^t = (a_{11}, \alpha, 0, \ldots, 0)^t, \tag{9.9}$$

donde α se seleccionará más adelante. Para simplificar la notación, si

$$\hat{\mathbf{w}} = (w_2, w_3, \ldots, w_n)^t \in \mathbb{R}^{n-1}, \quad \hat{\mathbf{y}} = (a_{21}, a_{31}, \ldots, a_{n1})^t \in \mathbb{R}^{n-1},$$

y $\hat{P}$ es una transformación de Householder $(n-1) \times (n-1)$

$$\hat{P} = I_{n-1} - 2\hat{\mathbf{w}}\hat{\mathbf{w}}^t.$$

Entonces, la ecuación (9.9) se convierte en

$$P^{(1)}\begin{bmatrix} a_{11} \\ a_{21} \\ a_{31} \\ \vdots \\ a_{n1} \end{bmatrix} = \left[\begin{array}{c|c} 1 & 0 \cdots\cdots 0 \\ \hline 0 & \\ \vdots & \hat{P} \\ 0 & \end{array}\right] \cdot \begin{bmatrix} a_{11} \\ \hline \hat{\mathbf{y}} \end{bmatrix} = \begin{bmatrix} a_{11} \\ \hline \hat{P}\hat{\mathbf{y}} \end{bmatrix} = \begin{bmatrix} a_{11} \\ \hline \alpha \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

con

$$\hat{P}\hat{\mathbf{y}} = (I_{n-1} - 2\hat{\mathbf{w}}\hat{\mathbf{w}}^t)\hat{\mathbf{y}} = \hat{\mathbf{y}} - 2(\hat{\mathbf{w}}^t\hat{\mathbf{y}})\hat{\mathbf{w}} = (\alpha, 0, \ldots, 0)^t. \tag{9.10}$$

Sea $r = \hat{\mathbf{w}}^t\hat{\mathbf{y}}$. Entonces

$$(\alpha, 0, \ldots, 0)^t = (a_{21} - 2rw_2, a_{31} - 2rw_3, \ldots, a_{n1} - 2rw_n)^t,$$

y podemos determinar todas las w_i una vez que conocemos α y r. Al equiparar los componentes da

$$\alpha = a_{21} - 2rw_2$$

y

$$0 = a_{j1} - 2rw_j, \quad \text{para cada } j = 3, \ldots, n.$$

Por lo tanto,

$$2rw_2 = a_{21} - \alpha \tag{9.11}$$

y

$$2rw_j = a_{j1}, \quad \text{para cada } j = 3, \ldots, n. \tag{9.12}$$

Al elevar al cuadrado ambos lados de cada una de las ecuaciones y sumar los términos correspondientes da

$$4r^2\sum_{j=2}^{n} w_j^2 = (a_{21} - \alpha)^2 + \sum_{j=3}^{n} a_{j1}^2.$$

Puesto que $\mathbf{w}^t\mathbf{w} = 1$ y $w_1 = 0$, tenemos $\sum_{j=2}^{n} w_j^2 = 1$,

$$4r^2 = \sum_{j=2}^{n} a_{j1}^2 - 2\alpha a_{21} + \alpha^2. \tag{9.13}$$

La ecuación (9.10) y el hecho de que P es ortogonal implica que

$$\alpha^2 = (\alpha, 0, \ldots, 0)(\alpha, 0, \ldots, 0)^t = (\hat{P}\hat{\mathbf{y}})^t\hat{P}\hat{\mathbf{y}} = \hat{\mathbf{y}}^t\hat{P}^t\hat{P}\hat{\mathbf{y}} = \hat{\mathbf{y}}^t\hat{\mathbf{y}}.$$

Por lo tanto,

$$\alpha^2 = \sum_{j=2}^{n} a_{j1}^2,$$

lo que, al sustituir en la ecuación (9.13), da

$$2r^2 = \sum_{j=2}^{n} a_{j1}^2 - \alpha a_{21}.$$

Para garantizar $2r^2 = 0$ si y sólo si $a_{21} = a_{31} = \cdots = a_{n1} = 0$, seleccionamos

$$\alpha = -\operatorname{sgn}(a_{21}) \left(\sum_{j=2}^{n} a_{j1}^2 \right)^{1/2},$$

lo cual implica que

$$2r^2 = \sum_{j=2}^{n} a_{j1}^2 + |a_{21}| \left(\sum_{j=2}^{n} a_{j1}^2 \right)^{1/2}.$$

Con esta selección de α y $2r^2$, resolvemos las ecuaciones (9.11) y (9.12) para obtener

$$w_2 = \frac{a_{21} - \alpha}{2r} \text{ y } w_j = \frac{a_{j1}}{2r}, \quad \text{para cada } j = 3, \ldots, n.$$

Para resumir la selección de $P^{(1)}$, tenemos

$$\alpha = -\operatorname{sgn}(a_{21}) \left(\sum_{j=2}^{n} a_{j1}^2 \right)^{1/2},$$

$$r = \left(\frac{1}{2}\alpha^2 - \frac{1}{2}a_{21}\alpha \right)^{1/2},$$

$$w_1 = 0,$$

$$w_2 = \frac{a_{21} - \alpha}{2r},$$

y

$$w_j = \frac{a_{j1}}{2r}, \text{ por cada } j = 3, \ldots, n.$$

Con esta selección

$$A^{(2)} = P^{(1)}AP^{(1)} = \begin{bmatrix} a_{11}^{(2)} & a_{12}^{(2)} & 0 & \cdots & 0 \\ a_{21}^{(2)} & a_{22}^{(2)} & a_{23}^{(2)} & \cdots & a_{2n}^{(2)} \\ 0 & a_{32}^{(2)} & a_{33}^{(2)} & \cdots & a_{3n}^{(2)} \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & a_{n2}^{(2)} & a_{n3}^{(2)} & \cdots & a_{nn}^{(2)} \end{bmatrix}.$$

Al haber encontrado $P^{(1)}$ y calculado $A^{(2)}$, el proceso se repite para $k = 2, 3, \ldots, n-2$ de acuerdo con lo siguiente:

$$\alpha = -\operatorname{sgn}(a_{k+1,k}^{(k)}) \left(\sum_{j=k+1}^{n} (a_{jk}^{(k)})^2 \right)^{1/2},$$

$$r = \left(\frac{1}{2}\alpha^2 - \frac{1}{2}\alpha\alpha_{k+1,k}^{(k)} \right)^{1/2},$$

$$w_1^{(k)} = w_2^{(k)} = \cdots = w_k^{(k)} = 0,$$

$$w_1^{(k)} = w_2^{(k)} = \cdots = w_k^{(k)} = 0,$$
$$w_{k+1}^{(k)} = \frac{a_{k+1,k}^{(k)} - \alpha}{2r},$$
$$w_j^{(k)} = \frac{a_{jk}^{(k)}}{2r}, \quad \text{para cada } j = k+2, k+3, \ldots, n,$$
$$P^{(k)} = I - 2\mathbf{w}^{(k)} \cdot (\mathbf{w}^{(k)})^t,$$

y

$$A^{(k+1)} = P^{(k)} A^{(k)} P^{(k)},$$

donde

$$A^{(k+1)} = \begin{bmatrix} a_{11}^{(k+1)} & a_{12}^{(k+1)} & 0 & \cdots & \cdots & \cdots & 0 \\ a_{21}^{(k+1)} & \ddots & \ddots & \ddots & & & \vdots \\ 0 & \ddots & \ddots & \ddots & \ddots & 0 & 0 \\ \vdots & & \ddots & a_{k+1,k}^{(k+1)} & a_{k+1,k+1}^{(k+1)} & a_{k+1,k+2}^{(k+1)} \cdots & a_{k+1,n}^{(k+1)} \\ \vdots & & & 0 & \vdots & \ddots & \vdots \\ 0 & \cdots & \cdots & 0 & a_{n,k+1}^{(k+1)} & \cdots & a_{nn}^{(k+1)} \end{bmatrix}.$$

Si continuamos de esta manera, se forma la matriz tridiagonal y simétrica $A^{(n-1)}$, donde

$$A^{(n-1)} = P^{(n-2)} P^{(n-3)} \cdots P^{(1)} A P^{(1)} \cdots P^{(n-3)} P^{(n-2)}.$$

Ejemplo 1 Aplique las transformaciones de Householder a la matriz simétrica 4×4

$$A = \begin{bmatrix} 4 & 1 & -2 & 2 \\ 1 & 2 & 0 & 1 \\ -2 & 0 & 3 & -2 \\ 2 & 1 & -2 & -1 \end{bmatrix}$$

para producir una matriz tridiagonal simétrica que es similar a A.

Solución Para la primera aplicación de una transformación de Householder,

$$\alpha = -(1)\left(\sum_{j=2}^{4} a_{j1}^2\right)^{1/2} = -3, \; r = \left(\frac{1}{2}(-3)^2 - \frac{1}{2}(1)(-3)\right)^{1/2} = \sqrt{6},$$
$$\mathbf{w} = \left(0, \frac{\sqrt{6}}{3}, -\frac{\sqrt{6}}{6}, \frac{\sqrt{6}}{6}\right),$$
$$P^{(1)} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} - 2\left(\frac{\sqrt{6}}{6}\right)^2 \begin{bmatrix} 0 \\ 2 \\ -1 \\ 1 \end{bmatrix} \cdot (0, 2, -1, 1)$$

$$= \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & -\frac{1}{3} & \frac{2}{3} & -\frac{2}{3} \\ 0 & \frac{2}{3} & \frac{2}{3} & \frac{1}{3} \\ 0 & -\frac{2}{3} & \frac{1}{3} & \frac{2}{3} \end{bmatrix},$$

y

$$A^{(2)} = \begin{bmatrix} 4 & -3 & 0 & 0 \\ -3 & \frac{10}{3} & 1 & \frac{4}{3} \\ 0 & 1 & \frac{5}{3} & -\frac{4}{3} \\ 0 & \frac{4}{3} & -\frac{4}{3} & -1 \end{bmatrix}.$$

Al continuar con la segunda iteración,

$$\alpha = -\frac{5}{3}, \quad r = \frac{2\sqrt{5}}{3}, \quad \mathbf{w} = \left(0, 0, 2\sqrt{5}, \frac{\sqrt{5}}{5}\right)^t,$$

$$P^{(2)} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & -\frac{3}{5} & -\frac{4}{5} \\ 0 & 0 & -\frac{4}{5} & \frac{3}{5} \end{bmatrix},$$ ■

y la matriz tridiagonal simétrica es

$$A^{(3)} = \begin{bmatrix} 4 & -3 & 0 & 0 \\ -3 & \frac{10}{3} & -\frac{5}{3} & 0 \\ 0 & -\frac{5}{3} & -\frac{33}{25} & \frac{68}{75} \\ 0 & 0 & \frac{68}{75} & \frac{149}{75} \end{bmatrix}.$$ ■

El algoritmo 9.5 realiza el método de Householder de acuerdo con lo que se describe aquí, a pesar de que se eluden las multiplicaciones reales de la matriz.

ALGORITMO 9.5

Método de Householder

Para obtener una matriz tridiagonal simétrica $A^{(n-1)}$ similar a la matriz simétrica $A = A^{(1)}$, construya las siguientes matrices
$A^{(2)}, A^{(3)}, \ldots, A^{(n-1)}$, donde $A^{(k)} = (a_{ij}^{(k)})$ para cada $k = 1, 2, \ldots, n-1$:

ENTRADA dimensión n; matriz A.

SALIDA $A^{(n-1)}$. (*En cada paso, A se puede sobreescribir.*)

Paso 1 Para $k = 1, 2, \ldots, n-2$ haga los pasos 2–14.

Paso 2 Determine

$$q = \sum_{j=k+1}^{n} \left(a_{jk}^{(k)}\right)^2.$$

Paso 3 Si $a_{k+1,k}^{(k)} = 0$ entonces determine $\alpha = -q^{1/2}$

si no determine $\alpha = -\dfrac{q^{1/2} a_{k+1,k}^{(k)}}{|a_{k+1,k}^{(k)}|}$.

Paso 4 Determine $RSQ = \alpha^2 - \alpha a_{k+1,k}^{(k)}$. (*Nota:* $RSQ = 2r^2$)

Paso 5 Determine $v_k = 0$; (*Nota:* $v_1 = \cdots = v_{k-1} = 0$, *pero no necesaria.*)
$v_{k+1} = a^{(k)}_{k+1,k} - \alpha$;
Para $j = k+2, \ldots, n$ determine $v_j = a^{(k)}_{jk}$.

$$\left(\textit{Nota: } \mathbf{w} = \left(\frac{1}{\sqrt{2RSQ}}\right)\mathbf{v} = \frac{1}{2r}\mathbf{v}.\right)$$

Paso 6 Para $j = k, k+1, \ldots, n$ determine $u_j = \left(\dfrac{1}{RSQ}\right)\displaystyle\sum_{i=k+1}^{n} a^{(k)}_{ji} v_i$.

$$\left(\textit{Nota: } \mathbf{u} = \left(\frac{1}{RSQ}\right) A^{(k)}\mathbf{v} = \frac{1}{2r^2}A^{(k)}\mathbf{v} = \frac{1}{r}A^{(k)}\mathbf{w}.\right)$$

Paso 7 Determine $PROD = \displaystyle\sum_{i=k+1}^{n} v_i u_i$.

$$\left(\textit{Nota: } PROD = \mathbf{v}^t\mathbf{u} = \frac{1}{2r^2}\mathbf{v}^t A^{(k)}\mathbf{v}.\right)$$

Paso 8 Para $j = k, k+1, \ldots, n$ determine $z_j = u_j - \left(\dfrac{PROD}{2RSQ}\right) v_j$.

$$\left(\textit{Nota: } \mathbf{z} = \mathbf{u} - \frac{1}{2RSQ}\mathbf{v}^t\mathbf{u}\mathbf{v} = \mathbf{u} - \frac{1}{4r^2}\mathbf{v}^t\mathbf{u}\mathbf{v}\right.$$

$$\left. = \mathbf{u} - \mathbf{w}\mathbf{w}^t\mathbf{u} = \frac{1}{r}A^{(k)}\mathbf{w} - \mathbf{w}\mathbf{w}^t\frac{1}{r}A^{(k)}\mathbf{w}.\right)$$

Paso 9 Para $l = k+1, k+2, \ldots, n-1$ haga los pasos 10 y 11.
(*Nota: Calcule* $A^{(k+1)} = A^{(k)} - \mathbf{v}\mathbf{z}^t - \mathbf{z}\mathbf{v}^t = (I - 2\mathbf{w}\mathbf{w}^t)A^{(k)}(I - 2\mathbf{w}\mathbf{w}^t)$.)

Paso 10 Para $j = l+1, \ldots, n$ determine
$a^{(k+1)}_{jl} = a^{(k)}_{jl} - v_l z_j - v_j z_l$;
$a^{(k+1)}_{lj} = a^{(k+1)}_{jl}$.

Paso 11 Determine $a^{(k+1)}_{ll} = a^{(k)}_{ll} - 2v_l z_l$.

Paso 12 Determine $a^{(k+1)}_{nn} = a^{(k)}_{nn} - 2v_n z_n$.

Paso 13 Para $j = k+2, \ldots, n$ determine $a^{(k+1)}_{kj} = a^{(k+1)}_{jk} = 0$.

Paso 14 Determine $a^{(k+1)}_{k+1,k} = a^{(k)}_{k+1,k} - v_{k+1} z_k$;
$a^{(k+1)}_{k,k+1} = a^{(k+1)}_{k+1,k}$.
(*Nota: Los otros elementos de* $A^{(k+1)}$ *son iguales a* $A^{(k)}$.)

Paso 15 SALIDA ($A^{(n-1)}$);
(*El proceso está completo.* $A^{(n-1)}$ *es simétrica, tridiagonal y similar a* A.)
PARE. ■

En la siguiente sección, examinaremos cómo se puede aplicar el algoritmo QR para determinar los eigenvalores de $A^{(n-1)}$, que son iguales a los de la matriz original A.

El algoritmo de Householder se puede aplicar a una matriz arbitraria $n \times n$, pero deben hacerse modificaciones para representar la posible falta de simetría. La matriz resultante $A^{(n-1)}$ no será tridiagonal a menos que la matriz original A sea simétrica, pero todas las entradas debajo de la subdiagonal inferior serán 0. Una matriz de este tipo recibe el nombre de *Hessenberg superior*. Es decir, $H = (h_{ij})$ es **Hessenberg superior** si $h_{ij} = 0$, para toda $i \geq j+2$.

Los siguientes pasos son las únicas modificaciones requeridas para matrices arbitrarias:

Paso 6 Para $j = 1, 2, \ldots, n$ determine $u_j = \dfrac{1}{RSQ} \displaystyle\sum_{i=k+1}^{n} a_{ji}^{(k)} v_i$;

$$y_j = \frac{1}{RSQ} \sum_{i=k+1}^{n} a_{ij}^{(k)} v_i.$$

Paso 8 Para $j = 1, 2, \ldots, n$ determine $z_j = u_j - \dfrac{PROD}{RSQ} v_j$.

Paso 9 Para $l = k+1, k+2, \ldots, n$ haga los pasos 10 y 11.

Paso 10 Para $j = 1, 2, \ldots, k$ determine $a_{jl}^{(k+1)} = a_{jl}^{(k)} - z_j v_l$;

$$a_{lj}^{(k+1)} = a_{lj}^{(k)} - y_j v_l.$$

Paso 11 Para $j = k+1, \ldots, n$ determine $a_{jl}^{(k+1)} = a_{jl}^{(k)} - z_j v_l - y_l v_j$.

Después de modificar estos pasos, borre del 12 al 14 y la salida $A^{(n-1)}$. Observe que el paso 7 permanece sin cambios.

La sección Conjunto de ejercicios 9.4 está disponible en línea. Encuentre la ruta de acceso en las páginas preliminares.

9.5 El algoritmo QR

En general, los métodos de deflación que se han analizado en la sección 9.3 no son adecuados para calcular todos los eigenvalores de una matriz debido al crecimiento del error de redondeo. En esta sección consideramos el algoritmo QR, una técnica de reducción de matriz que se usa para determinar en forma sistemática todos los eigenvalores de una matriz simétrica.

Para aplicar el método QR comenzamos con una matriz simétrica en forma diagonal; es decir, las únicas entradas diferentes de cero se encuentran ya sea sobre la diagonal o sobre las subdiagonales directamente sobre la diagonal o por debajo de ella. Si ésta no es la forma de una matriz simétrica, el primer paso es aplicar el método de Householder para calcular una matriz tridiagonal simétrica similar a la matriz dada.

En el resto de esta sección, se supondrá que la matriz simétrica para la que se calculan estos eigenvalores es tridiagonal. Si dejamos que A denote una matriz de este tipo, podemos simplificar la notación, en cierta medida, al etiquetar las entradas de A de acuerdo con lo siguiente:

$$A = \begin{bmatrix} a_1 & b_2 & 0 & \cdots & 0 \\ b_2 & a_2 & b_3 & \ddots & \vdots \\ 0 & b_3 & a_3 & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & b_n \\ 0 & \cdots & 0 & b_n & a_n \end{bmatrix}. \tag{9.14}$$

Si $b_2 = 0$ o $b_n = 0$, entonces la matriz 1×1 $[a_1]$ o $[a_n]$ produce de inmediato un eigenvalor a_1 o a_n de A. El método QR toma la ventaja de esta observación al disminuir sucesivamente los valores de las entradas por debajo de la diagonal principal hasta $b_2 \approx 0$ o $b_n \approx 0$.

Cuando $b_j = 0$ para algunas j, donde $2 < j < n$, es posible reducir el problema para considerar, en lugar de A, las matrices más pequeñas

$$\begin{bmatrix} a_1 & b_2 & 0 & \cdots & 0 \\ b_2 & a_2 & b_3 & \ddots & \vdots \\ 0 & b_3 & a_3 & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & b_{j-1} \\ 0 & \cdots & 0 & b_{j-1} & a_{j-1} \end{bmatrix} \quad \text{y} \quad \begin{bmatrix} a_j & b_{j+1} & 0 & \cdots & 0 \\ b_{j+1} & a_{j+1} & b_{j+2} & \ddots & \vdots \\ 0 & b_{j+2} & a_{j+2} & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & b_n \\ 0 & \cdots & 0 & b_n & a_n \end{bmatrix}. \tag{9.15}$$

Si ninguna b_j es cero, el método QR continúa al formar una sucesión de matrices $A = A^{(1)}, A^{(2)}, A^{(3)}, \ldots$, de acuerdo con lo siguiente:

i. $A^{(1)} = A$ se factoriza como producto de $A^{(1)} = Q^{(1)}R^{(1)}$, donde $Q^{(1)}$ es ortogonal y $R^{(1)}$ es triangular superior.

ii. $A^{(2)}$ se define como $A^{(2)} = R^{(1)}Q^{(1)}$.

En general, $A^{(i)}$ se factoriza como un producto $A^{(i)} = Q^{(i)}R^{(i)}$ de una matriz ortogonal $Q^{(i)}$ y una matriz triangular superior $R^{(i)}$. Entonces, $A^{(i+1)}$ se define mediante el producto de $R^{(i)}$ y $Q^{(i)}$ en dirección inversa $A^{(i+1)} = R^{(i)}Q^{(i)}$. Puesto que $Q^{(i)}$ es ortogonal, $R^{(i)} = Q^{(i)t}A^{(i)}$, y

$$A^{(i+1)} = R^{(i)}Q^{(i)} = (Q^{(i)t}A^{(i)})Q^{(i)} = Q^{(i)t}A^{(i)}Q^{(i)}. \tag{9.16}$$

Esto garantiza que $A^{(i+1)}$ es simétrica con los mismos eigenvalores que $A^{(i)}$. Por la manera en que definimos $Q^{(i)}$ y $R^{(i)}$, también garantizamos que $A^{(i+1)}$ es tridiagonal.

Al continuar mediante inducción, $A^{(i+1)}$ tiene los mismos eigenvalores que la matriz original A, y $A^{(i+1)}$ tiende a la matriz diagonal con los eigenvalores de A a lo largo de la diagonal.

Matrices de rotación

Para describir la construcción de las matrices de factorización $Q^{(i)}$ y $R^{(i)}$, necesitamos la noción de *matriz de rotación.*

Definición 9.23 Una **matriz de rotación** P difiere de la matriz identidad en máximo cuatro elementos. Estos cuatro elementos son de la forma

$$p_{ii} = p_{jj} = \cos\theta \quad \text{y} \quad p_{ij} = -p_{ji} = \operatorname{sen}\theta,$$

para algunos θ y algunas $i \neq j$. ■

Si A es la matriz de rotación 2×2

$$A = \begin{bmatrix} \cos\theta & -\operatorname{sen}\theta \\ \operatorname{sen}\theta & \cos\theta \end{bmatrix},$$

entonces $A\mathbf{x}$ rota a $\mathbf{x}$ en el sentido contrario a las manecillas del reloj un ángulo θ.

Es fácil mostrar (consulte el ejercicio 12) que, para cualquier matriz de rotación P, la matriz AP difiere de A sólo en la i-ésima y la j-ésima columnas y la matriz PA difiere de A sólo en la i-ésima y la j-ésima filas. Para cualquier $i \neq j$, el ángulo θ se puede seleccionar de tal forma que el producto PA tiene una entrada cero para $(PA)_{ij}$. Además, todas las matrices de rotación P son ortogonales porque la definición implica que $PP^t = I$.

Ejemplo 1 Encuentre una matriz de rotación P con la propiedad de que PA tiene una entrada cero en la segunda fila y la primera columna, donde

A menudo éstas reciben el nombre de rotaciones de Givens porque James Wallace Givens (1910-1993) las usó en la década de 1950 cuando estaba en los Laboratorios Nacionales Argonne.

$$A = \begin{bmatrix} 3 & 1 & 0 \\ 1 & 3 & 1 \\ 0 & 1 & 3 \end{bmatrix}.$$

Solución La forma de P es

$$P = \begin{bmatrix} \cos\theta & \operatorname{sen}\theta & 0 \\ -\operatorname{sen}\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix}, \text{ de modo que } PA = \begin{bmatrix} 3\cos\theta + \operatorname{sen}\theta & \cos\theta + 3\operatorname{sen}\theta & \operatorname{sen}\theta \\ -3\operatorname{sen}\theta + \cos\theta & -\operatorname{sen}\theta + 3\cos\theta & \cos\theta \\ 0 & 1 & 3 \end{bmatrix}.$$

El ángulo θ se selecciona de tal forma que $-3 \operatorname{sen} \theta + \cos \theta = 0$, es decir; de tal forma que $\theta = \dfrac{1}{3}$. Por lo tanto,

$$\cos \theta = \frac{3\sqrt{10}}{10}, \quad \operatorname{sen} \theta = \frac{\sqrt{10}}{10}$$

y

$$PA = \begin{bmatrix} \frac{3\sqrt{10}}{10} & \frac{\sqrt{10}}{10} & 0 \\ -\frac{\sqrt{10}}{10} & \frac{3\sqrt{10}}{10} & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 3 & 1 & 0 \\ 1 & 3 & 1 \\ 0 & 1 & 3 \end{bmatrix} = \begin{bmatrix} \sqrt{10} & \frac{3}{5}\sqrt{10} & \frac{1}{10}\sqrt{10} \\ 0 & \frac{4}{5}\sqrt{10} & \frac{3}{10}\sqrt{10} \\ 0 & 1 & 3 \end{bmatrix}.$$

Observe que la matriz resultante no es ni simétrica ni diagonal. ■

La factorización de $A^{(1)}$ como $A^{(1)} = Q^{(1)} R^{(1)}$ usa un producto de $n - 1$ matrices de rotación para construir

$$R^{(1)} = P_n P_{n-1} \cdots P_2 A^{(1)}.$$

Primero seleccionamos la matriz de rotación P_2 con

$$p_{11} = p_{22} = \cos \theta_2 \quad \text{y} \quad p_{12} = -p_{21} = \operatorname{sen} \theta_2,$$

donde

$$\operatorname{sen} \theta_2 = \frac{b_2}{\sqrt{b_2^2 + a_1^2}} \quad \text{y} \quad \cos \theta_2 = \frac{a_1}{\sqrt{b_2^2 + a_1^2}}.$$

Esta selección da

$$(-\operatorname{sen} \theta_2)a_1 + (\cos \theta_2)b_2 = \frac{-b_2 a_1}{\sqrt{b_2^2 + a_1^2}} + \frac{a_1 b_2}{\sqrt{b_2^2 + a_1^2}} = 0$$

para la entrada en la posición (2, 1), es decir, en la segunda fila y la primera columna del producto $P_2 A^{(1)}$. Por lo que la matriz

$$A_2^{(1)} = P_2 A^{(1)}$$

tiene un cero en la posición (2, 1).

La multiplicación $P_2 A^{(1)}$ afecta ambas filas, 1 y 2 de $A^{(1)}$, por lo que la matriz $A_2^{(1)}$ no necesariamente retiene cero entradas en las posiciones (1, 3) (1, 4), . . . , y (1, n). Sin embargo, $A^{(1)}$ es tridiagonal, por lo que (1, 4), . . . , (1, n) entradas de $A_2^{(1)}$ también deben ser 0. Sólo la entrada (1, 3), la de la primera fila y la tercera columna, se puede volver diferente a cero en $A_2^{(1)}$.

En general, la matriz P_k se selecciona de tal forma que la entrada $(k, k-1)$ en $A_k^{(1)} = P_k A_{k-1}^{(1)}$ es cero. Este resultado en la entrada $(k-1, k+1)$ se vuelve diferente de cero. La matriz $A_k^{(1)}$ tiene la forma

$$
A_k^{(1)} = \begin{bmatrix}
z_1 & q_1 & r_1 & 0 & \cdots & \cdots & \cdots & \cdots & 0 \\
0 & \ddots & \ddots & \ddots & \ddots & & & & \vdots \\
0 & \ddots & \ddots & \ddots & \ddots & \ddots & & & \vdots \\
\vdots & \ddots & 0 & z_{k-1} & q_{k-1} & r_{k-1} & \ddots & & \vdots \\
\vdots & & & 0 & x_k & y_k & 0 & \ddots & 0 \\
\vdots & & & & b_{k+1} & a_{k+1} & b_{k+2} & \ddots & 0 \\
\vdots & & & & & \ddots & \ddots & \ddots & 0 \\
\vdots & & & & & & \ddots & \ddots & b_n \\
0 & \cdots & \cdots & \cdots & \cdots & \cdots & 0 & b_n & a_n
\end{bmatrix},
$$

y P_{k+1} tiene la forma

$$
P_{k+1} = \left[\begin{array}{c|cc|c}
I_{k-1} & \multicolumn{2}{c|}{O} & O \\ \hline
 & c_{k+1} & s_{k+1} & \\
O & & & O \\
 & -s_{k+1} & c_{k+1} & \\ \hline
O & \multicolumn{2}{c|}{O} & I_{n-k-1}
\end{array}\right] \begin{array}{l} \\ \leftarrow \text{fila } k \\ \\ \\ \\ \end{array}, \qquad (9.17)
$$

$$
\qquad\qquad\qquad \uparrow \text{ columna } k
$$

donde 0 denota la matriz dimensional adecuada con todas las entradas cero.

Las constantes $c_{k+1} = \cos\theta_{k+1}$ y $s_{k+1} = \operatorname{sen}\theta_{k+1}$ en P_{k+1} se seleccionan de tal forma que la entrada $(k+1, k)$ en $A_{k+1}^{(1)}$ sea cero; es decir, $-s_{k+1}x_k + c_{k+1}b_{k+1} = 0$.

Puesto que $c_{k+1}^2 + s_{k+1}^2 = 1$, la solución de esta ecuación es

$$
s_{k+1} = \frac{b_{k+1}}{\sqrt{b_{k+1}^2 + x_k^2}} \quad \text{y} \quad c_{k+1} = \frac{x_k}{\sqrt{b_{k+1}^2 + x_k^2}},
$$

y $A_{k+1}^{(1)}$ tiene la forma

$$
A_{k+1}^{(1)} = \begin{bmatrix}
z_1 & q_1 & r_1 & 0 & \cdots & \cdots & \cdots & \cdots & \cdots \\
0 & \ddots & \ddots & \ddots & \ddots & & & & \vdots \\
0 & \ddots & \ddots & \ddots & \ddots & \ddots & & & \vdots \\
\vdots & \ddots & 0 & z_k & q_k & r_k & \ddots & & \vdots \\
\vdots & & & 0 & x_{k+1} & y_{k+1} & 0 & \ddots & \vdots \\
\vdots & & & & b_{k+2} & a_{k+2} & b_{k+3} & \ddots & 0 \\
\vdots & & & & & \ddots & \ddots & \ddots & 0 \\
\vdots & & & & & & \ddots & \ddots & b_n \\
0 & \cdots & \cdots & \cdots & \cdots & \cdots & 0 & b_n & a_n
\end{bmatrix}.
$$

Si continuamos con esta construcción en la sucesión $P_2, \ldots, P_n$, obtenemos la matriz triangular superior

$$R^{(1)} \equiv A_n^{(1)} = \begin{bmatrix} z_1 & q_1 & r_1 & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & \ddots & \ddots & r_{n-2} \\ \vdots & & & \ddots & z_{n-1} & q_{n-1} \\ 0 & \cdots & \cdots & \cdots & 0 & x_n \end{bmatrix}.$$

La otra mitad de la factorización QR es la matriz

$$Q^{(1)} = P_2^t P_3^t \cdots P_n^t$$

debido a la ortogonalidad de las matrices de rotación implica que

$$Q^{(1)}R^{(1)} = (P_2^t P_3^t \cdots P_n^t) \cdot (P_n \cdots P_3 P_2)A^{(1)} = A^{(1)}.$$

La matriz $Q^{(1)}$ es ortogonal porque

$$(Q^{(1)})^t Q^{(1)} = (P_2^t P_3^t \cdots P_n^t)^t (P_2^t P_3^t \cdots P_n^t) = (P_n \cdots P_3 P_2) \cdot (P_2^t P_3^t \cdots P_n^t) = I.$$

Además, $Q^{(1)}$ es una matriz Hessenberg superior. Para observar porqué esto es verdad, puede seguir los pasos en los ejercicios 13 y 14.

Por consiguiente, $A^{(2)} = R^{(1)}Q^{(1)}$ también es una matriz Hessenberg superior porque al multiplicar $Q^{(1)}$ a la izquierda de la matriz triangular superior $R^{(1)}$ no se afectan las entradas en el triángulo inferior. Ya sabemos que es simétrica, por lo que $A^{(2)}$ es tridiagonal.

En general, las entradas fuera de la diagonal de $A^{(2)}$, serán de magnitud más pequeña que las entradas correspondientes de $A^{(1)}$, por lo que $A^{(2)}$ está más cerca de ser una matriz diagonal que $A^{(1)}$. El proceso se repite para construir $A^{(3)}, A^{(4)}, \ldots$ hasta obtener convergencia satisfactoria.

Ejemplo 2 Aplique una iteración del método QR a la matriz dada en el ejemplo 1:

$$A = \begin{bmatrix} 3 & 1 & 0 \\ 1 & 3 & 1 \\ 0 & 1 & 3 \end{bmatrix}.$$

Solución Sea $A^{(1)} = A$ una matriz dada y P_2 representa la matriz de rotación determinada en el ejemplo 1. Encontramos, por medio de la notación presentada en el método QR, que

$$A_2^{(1)} = P_2 A^{(1)} = \begin{bmatrix} \frac{3\sqrt{10}}{10} & \frac{\sqrt{10}}{10} & 0 \\ -\frac{\sqrt{10}}{10} & \frac{3\sqrt{10}}{10} & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 3 & 1 & 0 \\ 1 & 3 & 1 \\ 0 & 1 & 3 \end{bmatrix} = \begin{bmatrix} \sqrt{10} & \frac{3}{5}\sqrt{10} & \frac{\sqrt{10}}{10} \\ 0 & \frac{4\sqrt{10}}{5} & \frac{3\sqrt{10}}{10} \\ 0 & 1 & 3 \end{bmatrix}$$

$$\equiv \begin{bmatrix} z_1 & q_1 & r_1 \\ 0 & x_2 & y_2 \\ 0 & b_3^{(1)} & a_3^{(1)} \end{bmatrix}.$$

Al continuar, tenemos

$$s_3 = \frac{b_3^{(1)}}{\sqrt{x_2^2 + (b_3^{(1)})^2}} = 0.36761 \quad \text{y} \quad c_3 = \frac{x_2}{\sqrt{x_2^2 + (b_3^{(1)})^2}} = 0.92998,$$

por lo que

$$R^{(1)} \equiv A_3^{(1)} = P_3 A_2^{(1)} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0.92998 & 0.36761 \\ 0 & -0.36761 & 0.92998 \end{bmatrix} \begin{bmatrix} \sqrt{10} & \frac{3}{5}\sqrt{10} & \frac{\sqrt{10}}{10} \\ 0 & \frac{4\sqrt{10}}{5} & \frac{3\sqrt{10}}{10} \\ 0 & 1 & 3 \end{bmatrix}$$

$$= \begin{bmatrix} \sqrt{10} & \frac{3}{5}\sqrt{10} & \frac{\sqrt{10}}{10} \\ 0 & 2.7203 & 1.9851 \\ 0 & 0 & 2.4412 \end{bmatrix}$$

y

$$Q^{(1)} = P_2^t P_3^t = \begin{bmatrix} \frac{3\sqrt{10}}{10} & -\frac{\sqrt{10}}{10} & 0 \\ \frac{\sqrt{10}}{10} & \frac{3\sqrt{10}}{10} & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0.92998 & -0.36761 \\ 0 & 0.36761 & 0.92998 \end{bmatrix}$$

$$= \begin{bmatrix} 0.94868 & -0.29409 & 0.11625 \\ 0.31623 & 0.88226 & -0.34874 \\ 0 & 0.36761 & 0.92998 \end{bmatrix}.$$

Por consiguiente,

$$A^{(2)} = R^{(1)} Q^{(1)} = \begin{bmatrix} \sqrt{10} & \frac{3}{5}\sqrt{10} & \frac{\sqrt{10}}{10} \\ 0 & 2.7203 & 1.9851 \\ 0 & 0 & 2.4412 \end{bmatrix} \begin{bmatrix} 0.94868 & -0.29409 & 0.11625 \\ 0.31623 & 0.88226 & -0.34874 \\ 0 & -0.36761 & 0.92998 \end{bmatrix}$$

$$= \begin{bmatrix} 3.6 & 0.86024 & 0 \\ 0.86024 & 3.12973 & 0.89740 \\ 0 & 0.89740 & 2.27027 \end{bmatrix}.$$

Los elementos diagonales de $A^{(2)}$ son aproximadamente 14% más pequeños que los de $A^{(1)}$, por lo que tenemos una reducción, pero no es considerable. Para disminuir por debajo de 0.001, necesitamos realizar 13 iteraciones del método QR. Al hacerlo, obtenemos

$$A^{(13)} = \begin{bmatrix} 4.4139 & 0.01941 & 0 \\ 0.01941 & 3.0003 & 0.00095 \\ 0 & 0.00095 & 1.5858 \end{bmatrix}.$$

Ésta daría un eigenvalor aproximado de 1.5858 y los eigenvalores restantes se podrían aproximar al considerar la matriz reducida

$$\begin{bmatrix} 4.4139 & 0.01941 \\ 0.01941 & 3.0003 \end{bmatrix}.$$

■

Aceleración de la convergencia

Si los eigenvalores de A tienen diferentes módulos con $|\lambda_1| > |\lambda_2| > \cdots > |\lambda_n|$, entonces la velocidad de convergencia de la entrada $b_{j+1}^{(i+1)}$ para 0 en la matriz $A^{(i+1)}$ depende de la razón $|\lambda_{j+1}/\lambda_j|$ (consulte [Fr]). La velocidad de convergencia de $b_{j+1}^{(i+1)}$ para 0 determina la velocidad a la que la entrada $a_j^{(i+1)}$ converge en el j-ésimo eigenvalor λ_j. Por lo tanto, la velocidad de convergencia puede ser lenta si $|\lambda_{j+1}/\lambda_j|$ no es significativamente menor que 1.

Para acelerar esta convergencia se usa una técnica de cambio similar a la que se usa con el método de potencia inversa en la sección 9.3. Se selecciona una constante σ cerca del eigenvalor de A. Esto modifica la factorización en la ecuación (9.16) para seleccionar $Q^{(i)}$ y $R^{(i)}$ de tal forma que

$$A^{(i)} - \sigma I = Q^{(i)} R^{(i)}, \tag{9.18}$$

y, por consiguiente, la matriz $A^{(i+1)}$ está definida como

$$A^{(i+1)} = R^{(i)}Q^{(i)} + \sigma I. \tag{9.19}$$

Con esta modificación, la velocidad de convergencia de $b_{j+1}^{(i+1)}$ para 0 depende de la razón $|(\lambda_{j+1} - \sigma)/(\lambda_j - \sigma)|$. Esto puede resultar en una mejora significativa sobre la velocidad original de convergencia de $a_j^{(i+1)}$ para λ_j si σ está cerca de λ_{j+1}, pero no cerca de λ_j.

Cambiamos σ en cada paso, de tal forma que cuando A tiene eigenvalores de diferentes módulos, $b_n^{(i+1)}$ converge a 0 más rápido que $b_j^{(i+1)}$ para cualquier entero j menor que n. Cuando $b_n^{(i+1)}$ es suficientemente pequeño, suponemos que $\lambda_n \approx a_n^{(i+1)}$, elimina las enésimas fila y columna de la matriz y continúa de la misma forma para encontrar una aproximación a λ_{n-1}. El proceso continúa hasta que se determina una aproximación para cada eigenvalor.

La técnica de cambio selecciona, en el i-ésimo paso, la constante de cambio σ_i, donde σ_i es el eigenvalor de la matriz

$$E^{(i)} = \begin{bmatrix} a_{n-1}^{(i)} & b_n^{(i)} \\ b_n^{(i)} & a_n^{(i)} \end{bmatrix}$$

que está más cerca de $a_n^{(i)}$. Este cambio traduce los eigenvalores de A mediante un factor σ_i. Con esta técnica de cambio, normalmente la convergencia es cúbica. (Consulte [WR], p. 270.) El método acumula estos cambios hasta que $b_n^{(i+1)} \approx 0$ y, después, añade cambios a $a_j^{(i+1)}$ para aproximar el eigenvalor λ_n.

Si A tiene eigenvalores del mismo módulo, $b_j^{(i+1)}$ quizá tienda a 0 para algunas $j \neq n$ a una velocidad más rápida que $b_n^{(i+1)}$. En este caso, la técnica de división de matriz descrita en (9.14) se puede usar para reducir el problema a uno que incluya un par de matrices de orden reducido.

Ejemplo 3 Incorpore el cambio al método QR para la matriz

$$A = \begin{bmatrix} 3 & 1 & 0 \\ 1 & 3 & 1 \\ 0 & 1 & 3 \end{bmatrix} = \begin{bmatrix} a_1^{(1)} & b_2^{(1)} & 0 \\ b_2^{(1)} & a_2^{(1)} & b_3^{(1)} \\ 0 & b_3^{(1)} & a_3^{(1)} \end{bmatrix}.$$

Solución Encontrar el parámetro de aceleración para el cambio requiere encontrar los eigenvalores de

$$\begin{bmatrix} a_2^{(1)} & b_3^{(1)} \\ b_3^{(1)} & a_3^{(1)} \end{bmatrix} = \begin{bmatrix} 3 & 1 \\ 1 & 3 \end{bmatrix},$$

que son $\mu_1 = 4$ y $\mu_2 = 2$. La selección del eigenvalor más cercano a $a_3^{(1)} = 3$ es arbitraria, y seleccionamos $\mu_2 = 2$ y cambiamos por esta cantidad. Entonces $\sigma_1 = 2$ y

$$\begin{bmatrix} d_1 & b_2^{(1)} & 0 \\ b_2^{(1)} & d_2 & b_3^{(1)} \\ 0 & b_3^{(1)} & d_3 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 1 \\ 0 & 1 & 1 \end{bmatrix}.$$

Si continuamos con los cambios obtenemos

$$x_1 = 1, \quad y_1 = 1, \quad z_1 = \sqrt{2}, \quad c_2 = \frac{\sqrt{2}}{2}, \quad s_2 = \frac{\sqrt{2}}{2},$$

$$q_1 = \sqrt{2}, \quad x_2 = 0, \quad r_1 = \frac{\sqrt{2}}{2}, \quad \text{y} \quad y_2 = \frac{\sqrt{2}}{2},$$

por lo que

$$A_2^{(1)} = \begin{bmatrix} \sqrt{2} & \sqrt{2} & \frac{\sqrt{2}}{2} \\ 0 & 0 & \sqrt{2} \\ 0 & 1 & 1 \end{bmatrix}.$$

Además,

$$z_2 = 1, \quad c_3 = 0, \quad s_3 = 1, \quad q_2 = 1, \quad \text{y} \quad x_3 = -\frac{\sqrt{2}}{2},$$

por lo que

$$R^{(1)} = A_3^{(1)} = \begin{bmatrix} \sqrt{2} & \sqrt{2} & \frac{\sqrt{2}}{2} \\ 0 & 1 & 1 \\ 0 & 0 & -\frac{\sqrt{2}}{2} \end{bmatrix}.$$

Para calcular $A^{(2)}$, tenemos

$$z_3 = -\frac{\sqrt{2}}{2}, \quad a_1^{(2)} = 2, \quad b_2^{(2)} = \frac{\sqrt{2}}{2}, \quad a_2^{(2)} = 1, \quad b_3^{(2)} = -\frac{\sqrt{2}}{2}, \quad \text{y} \quad a_3^{(2)} = 0,$$

por lo que

$$A^{(2)} = R^{(1)}Q^{(1)} = \begin{bmatrix} 2 & \frac{\sqrt{2}}{2} & 0 \\ \frac{\sqrt{2}}{2} & 1 & -\frac{\sqrt{2}}{2} \\ 0 & -\frac{\sqrt{2}}{2} & 0 \end{bmatrix}.$$

Una iteración del método QR está completa. Ni $b_2^{(2)} = \sqrt{2}/2$ ni $b_3^{(2)} = -\sqrt{2}/2$ son pequeñas, por lo que se realiza otra iteración del método QR. Para esta iteración, calculamos los eigenvalores $\frac{1}{2} \pm \frac{1}{2}\sqrt{3}$ de la matriz

$$\begin{bmatrix} a_2^{(2)} & b_3^{(2)} \\ b_3^{(2)} & a_3^{(2)} \end{bmatrix} = \begin{bmatrix} 1 & -\frac{\sqrt{2}}{2} \\ -\frac{\sqrt{2}}{2} & 0 \end{bmatrix}$$

y seleccionamos $\sigma_2 = \frac{1}{2} - \frac{1}{2}\sqrt{3}$, el eigenvalor más cercano a $a_3^{(2)} = 0$. Al completar los cálculos, obtenemos

$$A^{(3)} = \begin{bmatrix} 2.6720277 & 0.37597448 & 0 \\ 0.37597448 & 1.4736080 & 0.030396964 \\ 0 & 0.030396964 & -0.047559530 \end{bmatrix}.$$

Si $b_3^{(3)} = 0.030396964$ es suficientemente pequeño, entonces la aproximación del eigenvalor λ_3 es 1.5864151, la suma de $a_3{}^{(3)}$ y los cambios $\sigma_1 + \sigma_2 = 2 + (1 - \sqrt{3})/2$. Al borrar la tercera fila y columna, obtenemos

$$A^{(3)} = \begin{bmatrix} 2.6720277 & 0.37597448 \\ 0.37597448 & 1.4736080 \end{bmatrix},$$

que tiene eigenvalores $\mu_1 = 2.7802140$ y $\mu_2 = 1.3654218$. Al sumar los cambios obtenemos las aproximaciones

$$\lambda_1 \approx 4.4141886 \quad \text{y} \quad \lambda_2 \approx 2.9993964.$$

Los eigenvalores reales de la matriz A son 4.41420, 3.00000, y 1.58579, por lo que el método QR dio cuatro dígitos significativos de exactitud en sólo dos iteraciones. ■

El algoritmo 9.6 implementa el método QR.

ALGORITMO
9.6

Método QR

Para obtener los eigenvalores de la matriz simétrica, tridiagonal $n \times n$

$$A \equiv A_1 = \begin{bmatrix} a_1^{(1)} & b_2^{(1)} & 0 & \cdots & 0 \\ b_2^{(1)} & a_2^{(1)} & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & b_n^{(1)} \\ 0 & \cdots & 0 & b_n^{(1)} & a_n^{(1)} \end{bmatrix}$$

ENTRADA $n; a_1^{(1)}, \ldots, a_n^{(1)}, b_2^{(1)}, \ldots, b_n^{(1)}$; tolerancia TOL; número máximo de iteraciones M.

SALIDA eigenvalores de A, o la división recomendada de A o un mensaje que indica que el número máximo de iteraciones fue superado.

Paso 1 Determine $k = 1$;
$SHIFT = 0$. (*Cambio acumulado.*)

Paso 2 Mientras $k \leq M$ haga los pasos 3–19.
Prueba de los pasos 3–7 para éxito.)

Paso 3 Si $|b_n^{(k)}| \leq TOL$ entonces determine $\lambda = a_n^{(k)} + SHIFT$;
SALIDA (λ);
determine $n = n - 1$.

Paso 4 Si $|b_2^{(k)}| \leq TOL$ entonces determine $\lambda = a_1^{(k)} + SHIFT$;
SALIDA (λ);
determine $n = n - 1$;
$a_1^{(k)} = a_2^{(k)}$;
para $j = 2, \ldots, n$
determine $a_j^{(k)} = a_{j+1}^{(k)}$;
$b_j^{(k)} = b_{j+1}^{(k)}$.

Paso 5 Si $n = 0$ entonces
PARE.

Paso 6 Si $n = 1$ entonces
determine $\lambda = a_1^{(k)} + SHIFT$;
SALIDA (λ);
PARE.

Paso 7 Para $j = 3, \ldots, n - 1$
si $|b_j^{(k)}| \leq TOL$ entonces
SALIDA (‘dividido en’, $a_1^{(k)}, \ldots, a_{j-1}^{(k)}, b_2^{(k)}, \ldots, b_{j-1}^{(k)}$,
‘y’,
$a_j^{(k)}, \ldots, a_n^{(k)}, b_{j+1}^{(k)}, \ldots, b_n^{(k)}, SHIFT$);
PARE.

Paso 8 (*Calcule el cambio.*)
Determine $b = -(a_{n-1}^{(k)} + a_n^{(k)})$;
$c = a_n^{(k)} a_{n-1}^{(k)} - \left[b_n^{(k)}\right]^2$;
$d = (b^2 - 4c)^{1/2}$.

Paso 9 Si $b > 0$ entonces determine $\mu_1 = -2c/(b+d)$;
$\mu_2 = -(b+d)/2$
si no determine $\mu_1 = (d-b)/2$;
$\mu_2 = 2c/(d-b)$.

Paso 10 Si $n = 2$ entonces determine $\lambda_1 = \mu_1 + SHIFT$;
$\lambda_2 = \mu_2 + SHIFT$;
SALIDA (λ_1, λ_2);
PARE.

Paso 11 Escoja σ por lo que $|\sigma - a_n^{(k)}| = \text{mín}\{|\mu_1 - a_n^{(k)}|, |\mu_2 - a_n^{(k)}|\}$.

Paso 12 (*Acumule el cambio.*)
Determine $SHIFT = SHIFT + \sigma$.

Paso 13 (*Realice el cambio.*)
Para $j = 1, \ldots, n$, determine $d_j = a_j^{(k)} - \sigma$.

Paso 14 (*Los pasos* 14 *y* 15 *calculan* $R^{(k)}$.)
Determine $x_1 = d_1$;
$y_1 = b_2$.

Paso 15 Para $j = 2, \ldots, n$

determine $z_{j-1} = \left\{x_{j-1}^2 + \left[b_j^{(k)}\right]^2\right\}^{1/2}$;

$$c_j = \frac{x_{j-1}}{z_{j-1}};$$

$$\sigma_j = \frac{b_j^{(k)}}{z_{j-1}};$$

$q_{j-1} = c_j y_{j-1} + s_j d_j$;
$x_j = -\sigma_j y_{j-1} + c_j d_j$;
Si $j \neq n$ entonces determine $r_{j-1} = \sigma_j b_{j+1}^{(k)}$;
$y_j = c_j b_{j+1}^{(k)}$.
$\left(A_j^{(k)} = P_j A_{j-1}^{(k)}\right.$ *se acaba de calcular y* $\left.R^{(k)} = A_n^{(k)}.\right)$

Paso 16 (*Pasos* 16–18 *calcule* $A^{(k+1)}$.)
Determine $z_n = x_n$;
$a_1^{(k+1)} = \sigma_2 q_1 + c_2 z_1$;
$b_2^{(k+1)} = \sigma_2 z_2$.

Paso 17 Para $j = 2, 3, \ldots, n-1$
determine $a_j^{(k+1)} = \sigma_{j+1} q_j + c_j c_{j+1} z_j$;
$b_{j+1}^{(k+1)} = \sigma_{j+1} z_{j+1}$.

Paso 18 Determine $a_n^{(k+1)} = c_n z_n$.

Paso 19 Determine $k = k + 1$.

Paso 20 SALIDA ('el número máximo de iteraciones excedido');
(*El procedimiento no fue exitoso.*)
PARE. ■

Se puede usar un procedimiento similar para encontrar las aproximaciones para los eigenvalores de una matriz simétrica $n \times n$. Primero, la matriz se reduce a una matriz Hessenberg superior similar mediante el algoritmo de Hessenberg para matrices no simétricas, descrito al final de la sección 9.4.

El proceso de factorización QR asume la siguiente forma. Primero,

$$H \equiv H^{(1)} = Q^{(1)} R^{(1)}. \tag{9.20}$$

Entonces $H^{(2)}$ se define por medio de

$$H^{(2)} = R^{(1)} Q^{(1)} \tag{9.21}$$

y se factoriza en

$$H^{(2)} = Q^{(2)} R^{(2)}. \tag{9.22}$$

El método para factorizar continúa con el mismo objetivo que el algoritmo QR para matrices simétricas. Es decir, las matrices se seleccionan para introducir ceros en las entradas adecuadas de la matriz y se usa un procedimiento de cambio similar al del método QR. Sin embargo, el cambio es, en cierta medida, complicado para las matrices no simétricas ya que pueden presentarse eigenvalores complejos con el mismo módulo. El proceso de cambio modifica los cálculos en las ecuaciones (9.20), (9.21) y (9.22) para obtener el método QR doble

$$H^{(1)} - \sigma_1 I = Q^{(1)} R^{(1)}, \qquad H^{(2)} = R^{(1)} Q^{(1)} + \sigma_1 I,$$
$$H^{(2)} - \sigma_2 I = Q^{(2)} R^{(2)}, \qquad H^{(3)} = R^{(2)} Q^{(2)} + \sigma_2 I,$$

donde σ_1 y σ_2 son conjugados complejos y $H^{(1)}, H^{(2)}, \ldots$ son matrices Hessenberg superior.

Es posible encontrar una descripción completa del método QR en los trabajos de Wilkinson [Wil2]. Algoritmos y programas detallados para éste y otros métodos que se usan con mayor frecuencia se proporcionan en [WR]. Remitimos al lector hacia estos trabajos si el método que hemos analizado no arroja resultados satisfactorios.

James Hardy Wilkinson (1919–1986) es mejor conocido por su amplio trabajo sobre métodos numéricos para resolver sistemas de ecuaciones lineales y problemas de eigenvalor. También desarrolló la técnica de álgebra lineal numérica de análisis de error regresivo.

El método QR se puede realizar de forma que producirá los eigenvectores de una matriz, así como eigenvalores, pero el algoritmo 9.6 no se ha diseñado para lograr esto. Si se necesitan los eigenvectores de una matriz simétrica, así como los eigenvalores, sugerimos que se use ya sea el método de potencia inversa después de haber utilizado los algoritmos 9.5 y 9.6 o una de las técnicas más poderosas mencionadas en [WR].

La sección Conjunto de ejercicios 9.5 está disponible en línea. Encuentre la ruta de acceso en las páginas preliminares.

9.6 Descomposición en valores singulares

En esta sección, consideramos la factorización de una matriz general A $m \times n$ en lo que se conoce como descomposición en valores singulares. Esta factorización toma la forma

$$A = U S V^t,$$

donde U es una matriz ortogonal $m \times m$, V es una matriz ortogonal $n \times n$ y S es una matriz $m \times n$, cuyos elementos diferentes de cero se encuentran a lo largo de la diagonal principal. En esta sección supondremos que $m \geq n$, y, en muchas aplicaciones importantes, m es mucho más grande que n.

La descomposición en valores singulares tiene una historia bastante larga, los matemáticos la consideraron por primera vez a finales del siglo XIX. Sin embargo, sus aplicaciones importantes tuvieron que esperar hasta que la potencia computacional estuviera disponible

en la segunda mitad del siglo XX, cuando los algoritmos pudieron desarrollarse para su implementación eficiente. Éste fue el trabajo principal de Gene Golub (1932–2007) en una serie de artículos en las décadas de 1960 y 1970. (Consulte, en especial, [GK] y [GR].) Una historia bastante completa de la técnica puede encontrarse en el artículo de G. W. Stewart, que está disponible por medio de internet en la dirección provista en [Stew3].

Antes de continuar con la descomposición en valores singulares, necesitamos describir algunas propiedades de las matrices arbitrarias.

Definición 9.24 Sea A una matriz $m \times n$.

i) El **rango** de A, que se denota rango(A), es el número de filas linealmente independientes en A.

ii) La **nulidad** de A, que se denota nulidad(A), es $n -$ rango(A) y describe el conjunto más grande de vectores linealmente independientes $\mathbf{v}$ en $\mathbb{R}^n$ para el que $A\mathbf{v} = \mathbf{0}$. ■

El rango y la nulidad de una matriz son importantes para caracterizar la conducta de la matriz. Cuando la matriz es cuadrada, por ejemplo, es invertible si y sólo si su nulidad es 0 y su rango es igual al tamaño de la matriz.

El siguiente es uno de los teoremas básicos en álgebra lineal.

Teorema 9.25 El número de filas linealmente independientes de una matriz A $m \times n$ es el mismo número de columnas linealmente independientes de A. ■

Necesitamos considerar las dos matrices cuadradas relacionadas con la matriz A $m \times n$, concretamente, la matriz $A^t A$ $m \times n$ y la matriz AA^t.

Teorema 9.26 Sea A una matriz $m \times n$.

i) Las matrices $A^t A$ y AA^t son simétricas.

ii) Nulidad$(A) =$ Nulidad$(A^t A)$.

iii) Rango$(A) =$ Rango$(A^t A)$.

iv) Los eigenvalores de $A^t A$ son reales y no negativos.

v) Los eigenvalores diferentes de cero de AA^t son iguales a los eigenvalores de $A^t A$.

Demostración **i)** Puesto que $(A^t A)^t = A^t(A^t)^t = A^t A$, esta matriz es simétrica, y, de igual forma, lo es AA^t.

ii) Sea $\mathbf{v} \neq \mathbf{0}$ un vector con $A\mathbf{v} = \mathbf{0}$. Entonces

$$(A^t A)\mathbf{v} = A^t(A\mathbf{v}) = A^t\mathbf{0} = \mathbf{0}, \quad \text{por lo que} \quad \text{nulidad}(A) \leq \text{Nulidad}(A^t A).$$

Ahora suponga que $\mathbf{v}$ es un vector $n \times 1$ con $A^t A\mathbf{v} = \mathbf{0}$. Entonces,

$$0 = \mathbf{v}^t A^t A\mathbf{v} = (A\mathbf{v})^t A\mathbf{v} = ||A\mathbf{v}||_2^2, \quad \text{lo cual implica que} \quad A\mathbf{v} = \mathbf{0}.$$

Por lo tanto, Nulidad$(A^t A) \leq$ Nulidad(A). Por consiguiente, Nulidad$(A^t A) =$ Nulidad(A).

iii) Las matrices A y $A^t A$ tienen n columnas y sus nulidades concuerdan, por lo que

$$\text{Rango}(A) = n - \text{Nulidad}(A) = n - \text{Nulidad }(A^t A) = \text{Rango}(A^t A).$$

iv) Las matrices $A^t A$ y AA^t son simétricas mediante la parte i), por lo que el corolario 9.17 implica que sus eigenvalores son números reales. Suponga que $\mathbf{v}$ es un eigenvector de $A^t A$ con $||\mathbf{v}||_2 = 1$ asociado al eigenvalor λ. Entonces

$$0 \leq ||A\mathbf{v}||_2^2 = (A\mathbf{v})^t(A\mathbf{v}) = \mathbf{v}^t A^t A\mathbf{v} = \mathbf{v}^t(A^t A\mathbf{v}) = \mathbf{v}^t(\lambda\mathbf{v}) = \lambda\mathbf{v}^t\mathbf{v} = \lambda||\mathbf{v}||_2^2 = \lambda.$$

La demostración de que los eigenvalores de AA^t son no negativos sigue la demostración de la parte v).

v) Sea v un eigenvector correspondiente al eigenvalor diferente de cero de $A^t A$. Entonces

$$A^t A\mathbf{v} = \lambda\mathbf{v} \quad \text{implica que} \quad (AA^t)A\mathbf{v} = \lambda A\mathbf{v}.$$

Sea $A\mathbf{v} = \mathbf{0}$, entonces, $A^t A\mathbf{v} = A^t\mathbf{0} = \mathbf{0}$, lo cual contradice la suposición de que $\lambda \neq 0$. Por lo tanto, $A\mathbf{v}$ 0 y $A\mathbf{v}$ es un eigenvector de AA^t relacionado con λ el recíproco también sigue este argumento porque si λ es un eigenvalor diferente de cero de $AA^t = (A^t)^t A^t$, entonces λ también es un eigenvalor de $A^t(A^t)^t = A^t A$. ■

En la sección 5 del capítulo 6 observamos qué tan efectiva puede ser la factorización al resolver los sistemas lineales de la forma $A\mathbf{x} = \mathbf{b}$ cuando la matriz A se usa en forma repetida para variar $\mathbf{b}$. En esta sección consideraremos una técnica para factorizar una matriz general $m \times n$. Tiene aplicaciones en muchas áreas, incluyendo el ajuste de datos por mínimos cuadrados, la compresión de imágenes, el procesamiento de señal y la estadística.

Construcción de una descomposición en valores singulares para una matriz A $m \times n$

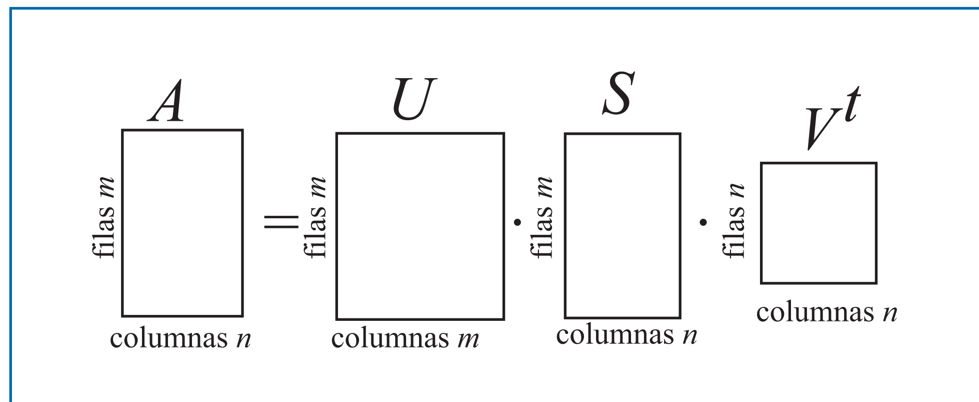
Una matriz no cuadrada A, es decir, con un número diferente de filas y columnas, no puede tener un eigenvalor porque $A\mathbf{x}$ y $\mathbf{x}$ serán vectores de diferentes tamaños. Sin embargo, existen números que desempeñan funciones para las matrices no cuadradas que son similares a las representadas por los eigenvalores para las matrices cuadradas. Una de las características más importantes de la descomposición en valores singulares de una matriz general es que permite una generalización de eigenvalores y eigenvectores en esta situación.

Nuestro objetivo es determinar una factorización de la matriz A $m \times n$, donde $m \geq n$, en la forma

$$A = U\, S\, V^t,$$

donde U es una matriz ortogonal $m \times m$, V es una matriz ortogonal $n \times n$ y S es una matriz diagonal $m \times n$; es decir, sólo entradas diferentes de cero $(S)_{ii} \equiv s_i \geq 0$, para $i = 1, \ldots, n$. (Consulte la figura 9.2.)

Figura 9.2



Construcción de S en la factorización $A = USV^t$

Construimos la matriz S al encontrar los eigenvalores de la matriz simétrica A^tA $n \times n$. Estos eigenvalores son todos números reales no negativos, los ordenamos de mayor a menor y los denotamos mediante

$$s_1^2 \geq s_2^2 \geq \cdots \geq s_k^2 > s_{k+1} = \cdots = s_n = 0.$$

Es decir, con S_k^2 denotamos el eigenvalor más pequeño diferente de cero de A^tA. Las raíces cuadradas positivas de estos eigenvalores de A^tA dan las entradas diagonales en S. Reciben el nombre de *valores singulares* de A. Por lo tanto.

$$S = \begin{bmatrix} s_1 & 0 & \cdots & 0 \\ 0 & s_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & s_n \\ 0 & \cdots & \cdots & 0 \\ \vdots & & & \vdots \\ 0 & \cdots & \cdots & 0 \end{bmatrix},$$

donde $s_i = 0$ cuando $k < i \leq n$.

Definición 9.27 Los **valores singulares** de una matriz A $m \times n$ son las raíces cuadradas positivas de los eigenvalores diferentes de cero de la matriz simétrica $A^t A$ $n \times n$. ■

Ejemplo 1 Determine los valores singulares de la matriz 5×3

$$A = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & 0 \\ 1 & 1 & 0 \end{bmatrix}.$$

Solución Tenemos

$$A^t = \begin{bmatrix} 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 0 & 0 \end{bmatrix}, \quad \text{por lo que} \quad A^t A = \begin{bmatrix} 2 & 1 & 1 \\ 1 & 4 & 1 \\ 1 & 1 & 2 \end{bmatrix}.$$

El polinomio característico de $A^t A$ es

$$\rho(A^t A) = \lambda^3 - 8\lambda^2 + 17\lambda - 10 = (\lambda - 5)(\lambda - 2)(\lambda - 1),$$

por lo que los eigenvalores de $A^t A$ son $\lambda_1 = s_1^2 = 5$, $\lambda_2 = s_2^2 = 2$, y $\lambda_3 = s_3^2 = 1$. Por consiguiente, los valores singulares de A son $s_1 = \sqrt{5}$, $s_2 = \sqrt{2}$, y $s_3 = 1$ y en la descomposición en valores singulares de A, tenemos

$$S = \begin{bmatrix} \sqrt{5} & 0 & 0 \\ 0 & \sqrt{2} & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

■

Cuando A es una matriz simétrica $n \times n$, todas las s_i^2 son eigenvalores de $A^2 = A^t A$, y éstos son cuadrados de los eigenvalores de A. (Consulte el ejercicio 17 de la sección 7.2.) Por lo que en este caso, los valores singulares son los valores absolutos de los eigenvalores de A. En el caso especial cuando A es definida positiva (o incluso definida no negativa), los eigenvalores y valores singulares de A son los mismos.

Construcción de V en la factorización $A = U\,S\,V^t$

La matriz A^tA $n \times n$ es simétrica, por lo que el teorema 9.16 en la sección 9.2 (consulte la página 430), tenemos una factorización

$$A^t A = V\,D\,V^t,$$

donde D es una matriz diagonal cuya diagonal consiste en eigenvalores de A^tA y V es una matriz ortogonal cuya i-ésima columna es un eigenvector con norma 1 l_2 asociado al eigenvalor de la i-ésima diagonal de D. La matriz diagonal específica depende del orden de los eigenvalores a lo largo de la diagonal. Seleccionamos D de tal forma que se escribe en orden decreciente. Las columnas, denotadas $\mathbf{v}_1^t, \mathbf{v}_2^t, \ldots, \mathbf{v}_n^t$, de la matriz ortogonal V $n \times n$ son eigenvectores ortonormales correspondientes a estos eigenvalores. Múltiples eigenvalores de A^tA permiten diversas selecciones de eigenvectores correspondientes, por lo que a pesar de que D está definida de manera única, la matriz V podría no estarlo. Sin embargo, no hay problema ya que podemos seleccionar cualquier V. Puesto que todos los eigenvalores de A^tA son no negativos, tenemos $d_{ii} = s_i^2$ para $1 \leq i \leq n$.

Construcción de U en la factorización $A = U\,S\,V^t$

Para construir la matriz U $m \times m$, primero consideramos los valores diferentes de cero $s_1 \geq s_2 \geq \cdots \geq s_k > 0$ y las columnas correspondientes en V determinadas por $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k$. Definimos

$$\mathbf{u}_i = \frac{1}{s_i} A\mathbf{v}_i, \quad \text{para } i = 1, 2, \ldots, k.$$

Las usamos como la primera k de las m columnas de U. Puesto que A es $m \times n$ y cada $\mathbf{v}_i$ es $n \times 1$, el vector $\mathbf{u}_i$ es $m \times 1$, según lo requerido. Además, para cada $1 \leq i \leq k$ y $1 \leq j \leq k$, el hecho es que los vectores $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ son eigenvectores de A^tA que forman un conjunto ortonormal implica que

$$\mathbf{u}_i^t\mathbf{u}_j = \left(\frac{1}{s_i}A\mathbf{v}_i\right)^t \frac{1}{s_j}A\mathbf{v}_j = \frac{1}{s_i s_j}\mathbf{v}_i^t A^t A\mathbf{v}_j = \frac{1}{s_i s_j}\mathbf{v}_i^t s_j^2 \mathbf{v}_j = \frac{s_j}{s_i}\mathbf{v}_i^t\mathbf{v}_j = \begin{cases} 0 & \text{si } i \neq j, \\ 1 & \text{si } i = j. \end{cases}$$

Por lo que, las primeras k columnas de U forman un conjunto ortonormal de vectores en $\mathbb{R}^m$. Sin embargo, necesitamos columnas adicionales $m - k$ de U. Para esto, primero necesitamos encontrar los vectores $m - k$ que, cuando se añadan a los vectores a partir de las primeras k columnas, nos darán un conjunto linealmente independiente. Entonces, podemos aplicar el proceso Gram-Schmidt para obtener las columnas adicionales adecuadas.

La matriz U no será única a menos que $k = m$ y entonces sólo si todos los eigenvalores de A^tA son únicos. La no singularidad no es preocupante; sólo necesitamos una matriz U de este tipo.

Verificación de la factorización $A = U\,S\,V^t$

Para verificar que este proceso en realidad da la factorización $A = USV^t$, primero recordemos que la transpuesta de una matriz ortogonal también es la inversa de la matriz (consulte la parte (i) de la sección 9.1 del teorema 9.10 en la página 428). Por lo tanto, para mostrar $A = USV^t$, podemos mostrar la declaración equivalente $AV = US$.

Los vectores $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ forman una base para $\mathbb{R}^n$, $A\mathbf{v}_i = s_i\mathbf{u}_i$, para $i = 1, \ldots, k$, y $A\mathbf{v}_i = \mathbf{0}$, para $i = k+1, \ldots, n$. Sólo las primeras k columnas de U producen entradas diferentes de cero en el producto US, por lo que tenemos

$$\begin{aligned}
AV &= A\,[\mathbf{v}_1\ \mathbf{v}_2\ \cdots \mathbf{v}_k\ \mathbf{v}_{k+1}\ \cdots \mathbf{v}_n] \\
&= [A\mathbf{v}_1\ A\mathbf{v}_2\ \cdots A\mathbf{v}_k\ A\mathbf{v}_{k+1}\ \cdots A\mathbf{v}_n] \\
&= [s_1\mathbf{u}_1\ s_2\mathbf{u}_2\ \cdots s_k\mathbf{u}_k\ \mathbf{0}\ \cdots \mathbf{0}] \\
&= [\mathbf{u}_1\ \mathbf{u}_2\ \cdots \mathbf{u}_k\ \mathbf{0}\ \cdots \mathbf{0}] \begin{bmatrix} s_1 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots & \vdots & & \vdots \\ \vdots & \ddots & \ddots & \vdots & \vdots & & \vdots \\ 0 & \cdots & 0 & s_k & 0 & \cdots & 0 \\ 0 & \cdots & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & & & \vdots & \vdots & & \vdots \\ 0 & \cdots & \cdots & 0 & 0 & \cdots & 0 \end{bmatrix} = US.
\end{aligned}$$

Esto completa la construcción de la descomposición en valores singulares de A.

Ejemplo 2 Determine la descomposición en valores singulares de la matriz 5×3

$$A = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & 0 \\ 1 & 1 & 0 \end{bmatrix}.$$

Solución En el ejemplo 1 encontramos que A tiene valores singulares $s_1 = \sqrt{5}$, $s_2 = \sqrt{2}$, y $s_3 = 1$, por lo que

$$S = \begin{bmatrix} \sqrt{5} & 0 & 0 \\ 0 & \sqrt{2} & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

Los eigenvectores de A^tA correspondientes a $s_1 = \sqrt{5}$, $s_2 = \sqrt{2}$, y $s_3 = 1$ son, respectivamente, $(1, 2, 1)^t$, $(1, -1, 1)^t$, y $(-1, 0, 1)^t$ (consulte el ejercicio 5). Al normalizar estos vectores y usar los valores para las columnas V obtenemos

$$V = \begin{bmatrix} \frac{\sqrt{6}}{6} & \frac{\sqrt{3}}{3} & -\frac{\sqrt{2}}{2} \\ \frac{\sqrt{6}}{3} & -\frac{\sqrt{3}}{3} & 0 \\ \frac{\sqrt{6}}{6} & \frac{\sqrt{3}}{3} & \frac{\sqrt{2}}{2} \end{bmatrix} \quad \text{y} \quad V^t = \begin{bmatrix} \frac{\sqrt{6}}{6} & \frac{\sqrt{6}}{3} & \frac{\sqrt{6}}{6} \\ \frac{\sqrt{3}}{3} & -\frac{\sqrt{3}}{3} & \frac{\sqrt{3}}{3} \\ -\frac{\sqrt{2}}{2} & 0 & \frac{\sqrt{2}}{2} \end{bmatrix}.$$

Las primeras tres columnas de U son, por lo tanto,

$$\mathbf{u}_1 = \frac{1}{\sqrt{5}} \cdot A\left(\frac{\sqrt{6}}{6}, \frac{\sqrt{6}}{3}, \frac{\sqrt{6}}{6}\right)^t = \left(\frac{\sqrt{30}}{15}, \frac{\sqrt{30}}{15}, \frac{\sqrt{30}}{10}, \frac{\sqrt{30}}{15}, \frac{\sqrt{30}}{10}\right)^t,$$

$$\mathbf{u}_2 = \frac{1}{\sqrt{2}} \cdot A\left(\frac{\sqrt{3}}{3}, -\frac{\sqrt{3}}{3}, \frac{\sqrt{3}}{3}\right)^t = \left(\frac{\sqrt{6}}{3}, -\frac{\sqrt{6}}{6}, 0, -\frac{\sqrt{6}}{6}, 0\right)^t, \text{ y}$$

$$\mathbf{u}_3 = 1 \cdot A\left(-\frac{\sqrt{2}}{2}, 0, \frac{\sqrt{2}}{2}\right)^t = \left(0, 0, \frac{\sqrt{2}}{2}, 0, -\frac{\sqrt{2}}{2}\right)^t.$$

Para determinar las dos columnas restantes de U primero necesitamos dos vectores $\mathbf{x}_4$ y $\mathbf{x}_5$ por lo que $\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3, \mathbf{x}_4, \mathbf{x}_5\}$ es un conjunto linealmente independiente. Entonces, aplicamos el proceso Gram-Schmidt para obtener $\mathbf{u}_4$ y $\mathbf{u}_5$ de tal forma que $\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3, \mathbf{u}_4, \mathbf{u}_5\}$ es un conjunto ortogonal. Dos vectores que satisfacen el requisito de independencia lineal y ortogonalidad son

$$\mathbf{u}_4 = (1, 1, -1, 1, -1)^t \quad \text{y} \quad \mathbf{u}_5 = (0, 1, 0, -1, 0)^t.$$

Al normalizar los vectores $\mathbf{u}_i$, para $i = 1, 2, 3, 4$ y 5 produce la matriz U y la descomposición en valores singulares como

$$A = U\,S\,V^t = \begin{bmatrix} \frac{\sqrt{30}}{15} & \frac{\sqrt{6}}{3} & 0 & \frac{\sqrt{5}}{5} & 0 \\ \frac{\sqrt{30}}{15} & -\frac{\sqrt{6}}{6} & 0 & \frac{\sqrt{5}}{5} & \frac{\sqrt{2}}{2} \\ \frac{\sqrt{30}}{10} & 0 & \frac{\sqrt{2}}{2} & -\frac{\sqrt{5}}{5} & 0 \\ \frac{\sqrt{30}}{15} & -\frac{\sqrt{6}}{6} & 0 & \frac{\sqrt{5}}{5} & -\frac{\sqrt{2}}{2} \\ \frac{\sqrt{30}}{10} & 0 & -\frac{\sqrt{2}}{2} & -\frac{\sqrt{5}}{5} & 0 \end{bmatrix} \begin{bmatrix} \sqrt{5} & 0 & 0 \\ 0 & \sqrt{2} & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

$$\times \begin{bmatrix} \frac{\sqrt{6}}{6} & \frac{\sqrt{6}}{3} & \frac{\sqrt{6}}{6} \\ \frac{\sqrt{3}}{3} & -\frac{\sqrt{3}}{3} & \frac{\sqrt{3}}{3} \\ -\frac{\sqrt{2}}{2} & 0 & \frac{\sqrt{2}}{2} \end{bmatrix}.$$

■

Una dificultad con el proceso en el ejemplo 2 es la necesidad de determinar los vectores adicionales $\mathbf{x}_4$ y $\mathbf{x}_5$ para proporcionar un conjunto linealmente independiente sobre el que podemos aplicar el proceso de Gram-Schmidt. Ahora consideraremos una forma de simplificarlo.

Un método alternativo para encontrar *U*

La parte v) del teorema 9.26 establece que los eigenvalores diferentes de cero de A^tA y los de AA^t son iguales. Además, los eigenvectores correspondientes de las matrices simétricas A^tA y AA^t forman subconjuntos ortonormales completos de $\mathbb{R}^n$ y $\mathbb{R}^m$, respectivamente. Así el conjunto ortonormal de n eigenvectores para A^tA forman las columnas de V, como se ha descrito antes y el conjunto ortonormal de m eigenvectores para AA^t forman las columnas de U de la misma forma.

En resumen, entonces, para determinar la descomposición en valores singulares de la matriz A $m \times n$, podemos:

- Encontrar los eigenvalores $s_1^2 \geq s_2^2 \geq \cdots \geq s_k^2 \geq s_{k+1} = \cdots = s_n = 0$ para la matriz simétrica A^tA y colocar la raíz cuadrada positiva de s_i^2 en la entrada $(S)_{ii}$ de la matriz S $m \times n$.
- Encontrar un conjunto de eigenvectores ortonormales $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}$ correspondiente a los eigenvalores de A^tA y, entonces, construir la matriz V $n \times n$ con estos vectores como columnas.
- Encontrar un conjunto de eigenvectores $\{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_m\}$ correspondientes a los eigenvalores de AA^t y construir la matriz U $m \times m$ con estos vectores como columnas.

Entonces A tiene la descomposición en valores singulares $A = U\,S\,V^t$.

Ejemplo 3 Determine la descomposición en valores singulares de la matriz 5×3

$$A = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & 0 \\ 1 & 1 & 0 \end{bmatrix}$$

al determinar U a partir de los eigenvectores de AA^t.

Solución Tenemos

$$AA^t = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & 0 \\ 1 & 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 2 & 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 1 & 1 \\ 1 & 1 & 2 & 1 & 1 \\ 0 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 2 \end{bmatrix},$$

que tiene los mismos eigenvalores diferentes de cero como $A^t A$, es decir, $\lambda_1 = 5$, $\lambda_2 = 2$, y $\lambda_3 = 1$ y, además, $\lambda_4 = 0$ y $\lambda_5 = 0$. Los eigenvectores correspondientes a estos eigenvalores son, respectivamente,

$$\mathbf{x}_1 = (2, 2, 3, 2, 3)^t, \quad \mathbf{x}_2 = (2, -1, 0, -1, 0)^t,$$
$$\mathbf{x}_3 = (0, 0, 1, 0, -1)^t, \quad \mathbf{x}_4 = (1, 2, -1, 0, -1)^t,$$

y $\mathbf{x}_5 = (0, 1, 0, -1, 0)^t$.

Ambos conjuntos $\{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4\}$ y $\{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_5\}$ son ortogonales porque son eigenvectores relacionados con distintos eigenvalores de la matriz simétrica AA^t. Sin embargo, $\mathbf{x}_4$ no es ortogonal a $\mathbf{x}_5$. Mantendremos $\mathbf{x}_4$ como uno de los eigenvectores que se usan para formar U y determinar el quinto vector que proporcionará el conjunto ortogonal. Para esto, usamos el proceso Gram-Schmidt como se describe en el teorema 9.8 en la página 427. Usando la notación en ese teorema, tenemos

$$\mathbf{v}_1 = \mathbf{x}_1, \mathbf{v}_2 = \mathbf{x}_2, \mathbf{v}_3 = \mathbf{x}_3, \mathbf{v}_4 = \mathbf{x}_4,$$

y, puesto que $\mathbf{x}_5$ es ortogonal para todo menos $\mathbf{x}_4$,

$$\begin{aligned} \mathbf{v}_5 = \mathbf{x}_5 - \frac{\mathbf{v}_4^t \mathbf{x}_5}{\mathbf{v}_4^t \mathbf{v}_4} \mathbf{v} \\ &= (0, 1, 0, -1, 0)^t - \frac{(1, 2, -1, 0, -1) \cdot (0, 1, 0, -1, 0)^t}{||(1, 2, -1, 0, -1)^t||_2^2}(1, 2, -1, 0, -1) \\ &= (0, 1, 0, -1, 0)^t - \frac{2}{7}(1, 2, -1, 0, -1)^t = -\frac{1}{7}(2, -3, -2, 7, -2)^t. \end{aligned}$$

Se verifica fácilmente que $\mathbf{v}_5$ es ortogonal a $\mathbf{v}_4 = \mathbf{x}_4$. También es ortogonal para los vectores en $\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\}$ porque es una combinación lineal de $\mathbf{x}_4$ y $\mathbf{x}_5$. Al normalizar estos vectores obtenemos la matriz U en la factorización. Por lo tanto,

$$U = [\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3, \mathbf{u}_4, \mathbf{u}_5] = \begin{bmatrix} \frac{\sqrt{30}}{15} & \frac{\sqrt{6}}{3} & 0 & \frac{\sqrt{7}}{7} & \frac{\sqrt{70}}{35} \\ \frac{\sqrt{30}}{15} & -\frac{\sqrt{6}}{6} & 0 & \frac{2\sqrt{7}}{7} & -\frac{3\sqrt{70}}{70} \\ \frac{\sqrt{30}}{10} & 0 & \frac{\sqrt{2}}{2} & -\frac{\sqrt{7}}{7} & -\frac{\sqrt{70}}{35} \\ \frac{\sqrt{30}}{15} & -\frac{\sqrt{6}}{6} & 0 & 0 & \frac{\sqrt{70}}{10} \\ \frac{\sqrt{30}}{10} & 0 & -\frac{\sqrt{2}}{2} & -\frac{\sqrt{7}}{7} & -\frac{\sqrt{70}}{35} \end{bmatrix}.$$

Esta es una U diferente a la que se encontró en el ejemplo 2, pero da una factorización válida $A = U\,S\,V^t$ usando las mismas S y V que en ese ejemplo. ■

Aproximación por mínimos cuadrados

La descomposición en valores singulares se aplica en muchas áreas, una de las cuales es un medio alterno para encontrar los polinomios de mínimos cuadrados para ajustar datos. Sea A una matriz $m \times n$, con $m > n$, y $\mathbf{b}$ un vector en $\mathbb{R}^m$. El objetivo de mínimos cuadrados es encontrar un vector $\mathbf{x}$ en $\mathbb{R}^n$ que minimizará $||A\mathbf{x} - \mathbf{b}||_2$.

Suponga que se conoce la descomposición en valores singulares de A, es decir,

$$A = U\,S\,V^t,$$

donde U es una matriz ortogonal $m \times m$, V es una matriz ortogonal $n \times n$ y S es una matriz $m \times n$ que contiene los valores singulares diferentes de cero en orden decreciente a lo largo de la diagonal principal en las primeras $k \leq n$ filas y entradas cero en otra parte. Puesto que tanto U como V son ortogonales, tenemos $U^{-1} = U^t$, $V^{-1} = V^t$, y mediante la parte iii) del teorema 9.10 en la sección 9.2 en la página 428, U y V preservan la norma l_2. Por consiguiente,

$$||A\mathbf{x} - \mathbf{b}||_2 = ||U\,S\,V^t\mathbf{x} - U\,U^t\mathbf{b}||_2 = ||S\,V^t\mathbf{x} - U^t\mathbf{b}||_2.$$

Sea $\mathbf{z} = V^t\mathbf{x}$ y $\mathbf{c} = U^t\mathbf{b}$. Entonces

$$\begin{aligned} ||A\mathbf{x} - \mathbf{b}||_2 &= ||(s_1z_1 - c_1, s_2z_2 - c_2, \ldots, s_kz_k - c_k, -c_{k+1}, \ldots, -c_m)^t||_2 \\ &= \left\{ \sum_{i=1}^{k} (s_iz_i - c_i)^2 + \sum_{i=k+1}^{m} (c_i)^2 \right\}^{1/2}. \end{aligned}$$

La norma se minimiza cuando se selecciona el vector $\mathbf{z}$ con

$$z_i = \begin{cases} \dfrac{c_i}{s_i}, & \text{cuando } i \leq k, \\ \text{arbitrario}, & \text{cuando } k < i \leq n. \end{cases}$$

Puesto que tanto $\mathbf{c} = U^t\mathbf{b}$ como $\mathbf{x} = V\mathbf{z}$ son fáciles de calcular, la solución de mínimos cuadrados también es fácil de encontrar.

Ejemplo 4 Use la técnica de descomposición en valores singulares para determinar el polinomio de mínimos cuadrados de grado 2 para los datos provistos en la tabla 9.5.

Tabla 9.5

i	x_i	y_i
1	0	1.0000
2	0.25	1.2840
3	0.50	1.6487
4	0.75	2.1170
5	1.00	2.7183

Solución Este problema se resolvió usando ecuaciones normales, como en el ejemplo 2 en la sección 8.1. Primero necesitamos determinar la forma adecuada para A, $\mathbf{x}$ y $\mathbf{b}$. En el ejemplo 2 en la sección 8.1, el problema se describió como encontrar a_0, a_1 y a_2 con

$$P_2(x) = a_0 + a_1x + a_2x^2.$$

A fin de expresar esto en forma de matriz, hacemos

$$\mathbf{x} = \begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} y_0 \\ y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} = \begin{bmatrix} 1.0000 \\ 1.2840 \\ 1.6487 \\ 2.1170 \\ 2.7183 \end{bmatrix}, \quad \text{y}$$

$$A = \begin{bmatrix} 1 & x_0 & x_0^2 \\ 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ 1 & x_3 & x_3^2 \\ 1 & x_4 & x_4^2 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0.25 & 0.0625 \\ 1 & 0.5 & 0.25 \\ 1 & 0.75 & 0.5625 \\ 1 & 1 & 1 \end{bmatrix}.$$

La descomposición en valores singulares de A tiene la forma $A = U\,S\,V^t$, donde

$$U = \begin{bmatrix} -0.2945 & -0.6327 & 0.6314 & -0.0143 & -0.3378 \\ -0.3466 & -0.4550 & -0.2104 & 0.2555 & 0.7505 \\ -0.4159 & -0.1942 & -0.5244 & -0.6809 & -0.2250 \\ -0.5025 & 0.1497 & -0.3107 & 0.6524 & -0.4505 \\ -0.6063 & 0.5767 & 0.4308 & -0.2127 & 0.2628 \end{bmatrix},$$

$$S = \begin{bmatrix} 2.7117 & 0 & 0 \\ 0 & 0.9371 & 0 \\ 0 & 0 & 0.1627 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad \text{y} \quad V^t = \begin{bmatrix} -0.7987 & -0.4712 & -0.3742 \\ -0.5929 & 0.5102 & 0.6231 \\ 0.1027 & -0.7195 & 0.6869 \end{bmatrix}.$$

Por lo que,

$$\mathbf{c} = U^t \begin{bmatrix} y_0 \\ y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} = \begin{bmatrix} -0.2945 & -0.6327 & 0.6314 & -0.0143 & -0.3378 \\ -0.3466 & -0.4550 & -0.2104 & 0.2555 & 0.7505 \\ -0.4159 & -0.1942 & -0.5244 & -0.6809 & -0.2250 \\ -0.5025 & 0.1497 & -0.3107 & 0.6524 & -0.4505 \\ -0.6063 & 0.5767 & 0.4308 & -0.2127 & 0.2628 \end{bmatrix}^t \begin{bmatrix} 1 \\ 1.284 \\ 1.6487 \\ 2.117 \\ 2.7183 \end{bmatrix}$$

$$= \begin{bmatrix} -4.1372 \\ 0.3473 \\ 0.0099 \\ -0.0059 \\ 0.0155 \end{bmatrix},$$

y los componentes de $\mathbf{z}$ son

$$z_1 = \frac{c_1}{s_1} = \frac{-4.1372}{2.7117} = -1.526, \quad z_2 = \frac{c_2}{s_2} = \frac{0.3473}{0.9371} = 0.3706, \quad \text{y}$$

$$z_3 = \frac{c_3}{s_3} = \frac{0.0099}{0.1627} = 0.0609.$$

Esto da los coeficientes de mínimos cuadrados para $P_2(x)$ como

$$\begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix} = \mathbf{x} = V\,\mathbf{z} = \begin{bmatrix} -0.7987 & -0.5929 & 0.1027 \\ -0.4712 & 0.5102 & -0.7195 \\ -0.3742 & 0.6231 & 0.6869 \end{bmatrix} \begin{bmatrix} -1.526 \\ 0.3706 \\ 0.0609 \end{bmatrix} = \begin{bmatrix} 1.005 \\ 0.8642 \\ 0.8437 \end{bmatrix},$$

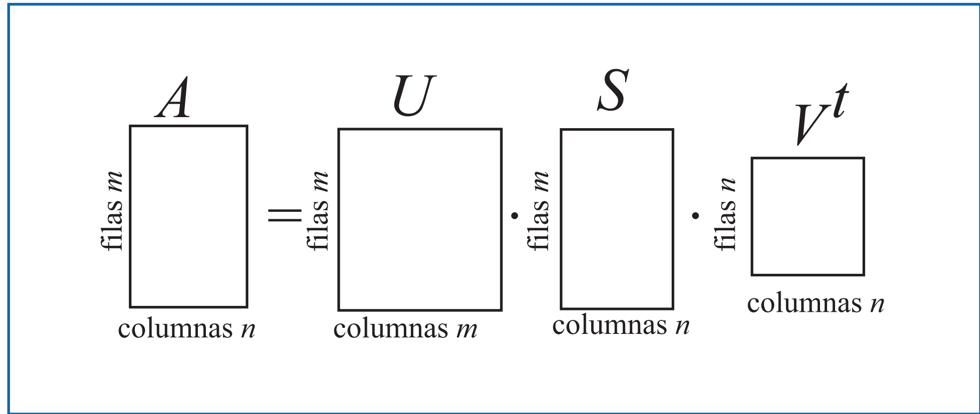
lo cual concuerda con los resultados en el ejemplo 2 de la sección 8.1. El error de mínimos cuadrados usando estos valores utiliza por lo menos dos componentes de $\mathbf{c}$ y es

$$||A\mathbf{x} - \mathbf{b}||_2 = \sqrt{c_4^2 + c_5^2} = \sqrt{(-0.0059)^2 + (0.0155)^2} = 0.0165. \quad ■$$

Otras aplicaciones

La razón de la importancia de la descomposición en valores singulares en muchas aplicaciones es que nos permite obtener las características más importantes de una matriz $m \times n$ usando una matriz que, a menudo, es de tamaño significativamente más pequeño. Puesto que los valores singulares están en la diagonal de S en orden decreciente, retener solamente las filas y columnas k de S produce la mejor aproximación posible para la matriz A. Como ilustración, consulte la figura 9.3, que indica la descomposición en valores singulares de la matriz A $n \times n$.

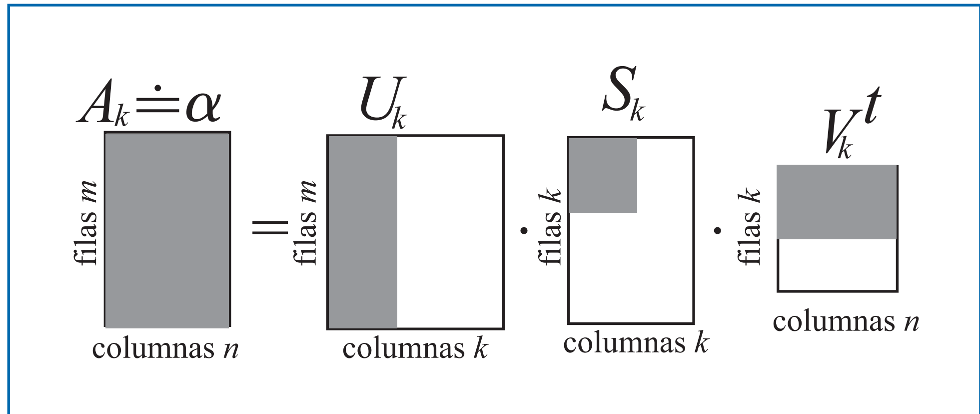
Figura 9.3



Reemplace la matriz S $n \times n$ con la matriz S_k $k \times k$ que contiene los valores singulares más significativos. Con toda certeza, éstos serán solamente los diferentes de cero, pero también podríamos eliminar algunos valores singulares que son relativamente pequeños.

Determine las matrices U_k y V_k^t, correspondientes $k \times n$ y $m \times k$ respectivamente, de acuerdo con el procedimiento de descomposición en valores singulares. Ésto se muestra sombreado en la figura 9.4.

Figura 9.4



Entonces la nueva matriz $A_k = U_k S_k V_k^t$ sigue siendo de tamaño $m \times n$ y requeriría $m \cdot n$ registros de almacenamiento para su representación. Sin embargo, en forma factorizada, el requisito de almacenamiento para los datos es $m \cdot k$ para U_k, k para S_k, y $n \cdot k$ para V_k^t para un total de $k(m + n + 1)$.

Suponga, por ejemplo, que $m = 2n$ y $k = n/3$. Entonces la matriz original A contiene $mn = 2n^2$ puntos de datos. La factorización que produce A_k, sin embargo, contiene solamente $mk = 2n^2/3$ para U_k, k para S_k, y $nk = n^2/3$ para V_k^t puntos de datos, que ocupan un total de $(n/3)(3n^2+1)$ registros de almacenamiento. Ésta es una reducción de aproximadamente 50% a partir de la cantidad requerida para almacenar toda la matriz A y los resultados en lo que se conoce como *compresión de datos*.

Ilustración En el ejemplo 2, demostramos que

$$A = U\,S\,V^t = \begin{bmatrix} \frac{\sqrt{30}}{15} & \frac{\sqrt{6}}{3} & 0 & \frac{\sqrt{5}}{5} & 0 \\ \frac{\sqrt{30}}{15} & -\frac{\sqrt{6}}{6} & 0 & \frac{\sqrt{5}}{5} & \frac{\sqrt{2}}{2} \\ \frac{\sqrt{30}}{10} & 0 & \frac{\sqrt{2}}{2} & -\frac{\sqrt{5}}{5} & 0 \\ \frac{\sqrt{30}}{15} & -\frac{\sqrt{6}}{6} & 0 & \frac{\sqrt{5}}{5} & -\frac{\sqrt{2}}{2} \\ \frac{\sqrt{30}}{10} & 0 & -\frac{\sqrt{2}}{2} & -\frac{\sqrt{5}}{5} & 0 \end{bmatrix} \begin{bmatrix} \sqrt{5} & 0 & 0 \\ 0 & \sqrt{2} & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \frac{\sqrt{6}}{6} & \frac{\sqrt{6}}{3} & \frac{\sqrt{6}}{6} \\ \frac{\sqrt{3}}{3} & -\frac{\sqrt{3}}{3} & \frac{\sqrt{3}}{3} \\ -\frac{\sqrt{2}}{2} & 0 & \frac{\sqrt{2}}{2} \end{bmatrix}.$$

Considere las matrices reducidas relacionadas con esta factorización

$$U_3 = \begin{bmatrix} \frac{\sqrt{30}}{15} & \frac{\sqrt{6}}{3} & 0 \\ \frac{\sqrt{30}}{15} & -\frac{\sqrt{6}}{6} & 0 \\ \frac{\sqrt{30}}{10} & 0 & \frac{\sqrt{2}}{2} \\ \frac{\sqrt{30}}{15} & -\frac{\sqrt{6}}{6} & 0 \\ \frac{\sqrt{30}}{10} & 0 & -\frac{\sqrt{2}}{2} \end{bmatrix}, \quad S_3 = \begin{bmatrix} \sqrt{5} & 0 & 0 \\ 0 & \sqrt{2} & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad \text{y}$$

$$V_3^t = \begin{bmatrix} \frac{\sqrt{6}}{6} & \frac{\sqrt{6}}{3} & \frac{\sqrt{6}}{6} \\ \frac{\sqrt{3}}{3} & -\frac{\sqrt{3}}{3} & \frac{\sqrt{3}}{3} \\ -\frac{\sqrt{2}}{2} & 0 & \frac{\sqrt{2}}{2} \end{bmatrix}.$$

Entonces

$$S_3 V_3^t = \begin{bmatrix} \frac{\sqrt{30}}{6} & \frac{\sqrt{30}}{3} & \frac{\sqrt{30}}{6} \\ \frac{\sqrt{6}}{3} & -\frac{\sqrt{6}}{3} & \frac{\sqrt{6}}{3} \\ -\frac{\sqrt{2}}{2} & 0 & \frac{\sqrt{2}}{2} \end{bmatrix} \quad \text{y} \quad A_3 = U_3 S_3 V_3^t = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & 0 \\ 1 & 1 & 0 \end{bmatrix}.$$ ■

Puesto que los cálculos en la ilustración se realizaron usando aritmética exacta, la matriz A_3 concuerda precisamente con la matriz original A. En general, se usaría la aritmética de dígitos finitos para hacer los cálculos y no se esperaría una concordancia absoluta. La esperanza es que la compresión de datos no resulte en una matriz A_k que difiera significativamente de la matriz original A y esto depende de las magnitudes relativas de los valores singulares de A. Cuando el rango de la matriz A es k, no habrá deterioro ya que sólo existen k filas de la matriz original A que son linealmente independientes y, en teoría, la matriz podría reducirse en una que tiene todos los ceros en sus últimas filas $m - k$ o columnas $n - k$. Cuando k es menor al rango de A, A_k diferirá de A, pero no siempre será en su detrimento.

Considere la situación que se presenta cuando A es una matriz que consiste en pixeles en una fotografía en escala de grises, que tal vez se tomó a gran distancia, como una fotografía de satélite de una parte de la Tierra. Es probable que la fotografía incluya *ruido*, es decir, datos que no representan verdaderamente la imagen, sino el deterioro de ésta mediante partículas atmosféricas, la calidad de las lentes y procesos de reproducción, y así sucesivamente. Los datos de ruido se incorporan en los datos determinados en A, pero con suerte este ruido es mucho menos significativo que la verdadera imagen. Esperamos que los valores singulares más grandes representen la verdadera imagen y que los más pequeños, los más cercanos a cero, sean las contribuciones del ruido. Al realizar la descomposición en valores singulares que solamente retiene esos valores singulares por encima de cierto umbral, podríamos ser capaces de eliminar la mayor parte del ruido y, en realidad, obtener una imagen que no sólo sea de menor tamaño sino también una representación verdadera de la fotografía original. (Consulte [AP] para más detalles, especialmente, la figura 3).

Otras aplicaciones importantes de la descomposición en valores singulares incluyen determinar números de condición efectivos para las matrices cuadradas (consulte el ejercicio 19), determinar el rango efectivo de una matriz y eliminar el ruido de la señal. Para más información sobre este importante tema y una interpretación geométrica de la factorización, examine el artículo de Kalman [Ka] y las referencias en ese documento. Para un estudio más completo y extenso de la teoría, consulte Golub y van Loan [GV].

La sección Conjunto de ejercicios 9.6 está disponible en línea. Encuentre la ruta de acceso en las páginas preliminares.

9.7 Software numérico

Las subrutinas en las bibliotecas IMSL y NAG, así como las subrutinas en Netlib y los comandos en MATLAB, Maple y Mathematica, están basadas en las incluidas en los paquetes EISPACK y LAPACK, que se analizan en la sección 1.4. En general, las subrutinas transforman una matriz en la forma adecuada para el método QR o una de sus modificaciones, como el método QL. Las subrutinas aproximan todos los eigenvalores y pueden aproximar un eigenvector relacionado para cada eigenvalor. En general, las matrices no simétricas se equilibran de tal forma que las sumas de las magnitudes de las entradas en cada fila y columna son casi iguales. Entonces se aplica el método Householder para determinar una matriz Hessenberg superior similar. Los eigenvalores se pueden calcular con el método QR o QL. También es posible aplicar la forma de Schur $S\,DS^t$, donde S es ortogonal y la diagonal de D mantiene los eigenvalores de A. Los eigenvectores correspondientes pueden entonces estar determinados. Para una matriz simétrica se calcula una matriz tridiagonal similar. Entonces, se pueden calcular los eigenvectores correspondientes usando el método QR o QL.

Existen rutinas especiales para encontrar todos los eigenvalores con un intervalo o una región, o que sólo encuentran el eigenvalor más grande o más pequeño. También existen subrutinas para determinar la precisión de la aproximación del eigenvalor y la sensibilidad del proceso al error de redondeo.

Un procedimiento MATLAB que calcula un número seleccionado de eigenvalores y eigenvectores está basado en el método Arnoldi implícitamente reiniciado de Sorensen [So]. Existe un paquete de software incluido en Netlib para resolver problemas de eigenvalores de gran dispersión que también está basado en el método Arnoldi implícitamente reiniciado. El método Arnoldi implícitamente reiniciado es un método de subespacio Krylov que encuentra una sucesión de subespacios Krylov que convergen en un subespacio que contiene los eigenvalores.

Las secciones Preguntas de análisis, Conceptos clave y Revisión del capítulo están disponibles en línea. Encuentre la ruta de acceso en las páginas preliminares.

CAPÍTULO

10

Soluciones numéricas de sistemas de ecuaciones no lineales

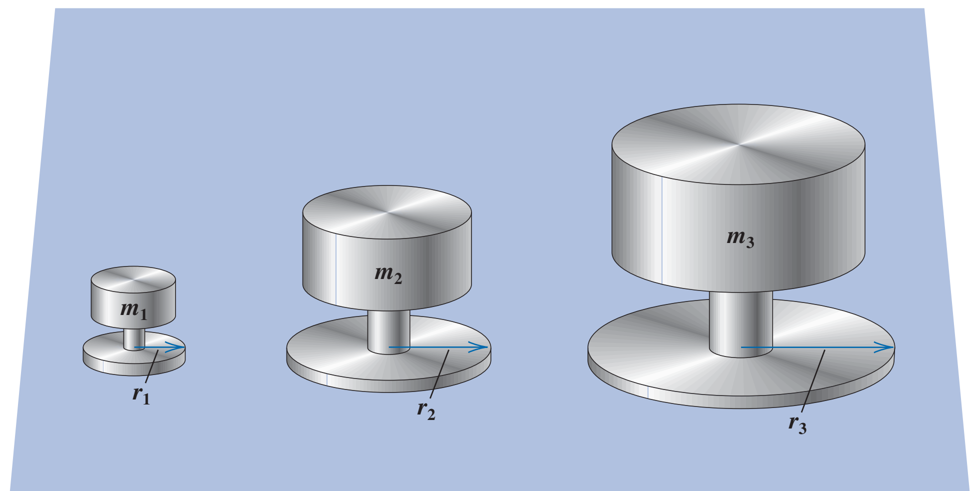
Introducción

La cantidad de presión requerida para hundir un objeto grande y pesado en un suelo homogéneo y blando que se encuentra sobre tierra de base dura se puede predecir por medio de la cantidad de presión requerida para hundir objetos más pequeños en la misma tierra. Especialmente, la cantidad de presión p para hundir una placa circular de radio r a una distancia d en tierra blanda, donde la tierra de base dura se encuentra a una distancia $D > d$ por debajo de la superficie, se puede aproximar mediante una ecuación de la forma

$$p = k_1 e^{k_2 r} + k_3 r,$$

donde k_1, k_2 y k_3 son constantes, dependiendo de d y de la consistencia del suelo pero no del radio de la placa.

Existen tres constantes desconocidas en esta ecuación, por lo que tres placas pequeñas con diferentes radios se hunden a la misma distancia. Esto determinará el tamaño mínimo de la placa requerido para mantener una carga grande. Se registran las cargas necesarias para este hundimiento, como se muestra en la figura adjunta.



Esto produce las tres ecuaciones no lineales

$$m_1 = k_1 e^{k_2 r_1} + k_3 r_1,$$

$$m_2 = k_1 e^{k_2 r_2} + k_3 r_2 \quad \text{y}$$

$$m_3 = k_1 e^{k_2 r_3} + k_3 r_3.$$

en los tres valores desconocidos k_1, k_2 y k_3. Normalmente, los métodos de aproximación numérica son necesarios para resolver sistemas de ecuaciones cuando éstas no son lineales. El ejercicio 10 de la sección 10.2 se preocupa por una aplicación del tipo descrito aquí.

La resolución de un sistema de ecuaciones no lineales es un problema que se evita siempre que sea posible, por regla general, al aproximar el sistema no lineal mediante un sistema de ecuaciones lineales. Cuando esto no es satisfactorio, el problema debe abordarse de manera directa. El enfoque más sencillo es adaptar los métodos del capítulo 2, los cuales aproximan las soluciones de una ecuación no lineal individual en una variable, que se aplicarán cuando el problema de una sola variable sea reemplazado por un problema vectorial que incluye todas las variables.

La principal herramienta en el capítulo 2 era el método de Newton, una técnica que, en general, es cuadráticamente convergente. Ésta es la primera técnica que modificamos para resolver los sistemas de ecuaciones no lineales. La aplicación del método de Newton, de acuerdo con lo modificado para los sistemas de ecuaciones, es bastante costosa y en la sección 10.3 describimos cómo se puede usar el método de secante para obtener aproximaciones con mayor facilidad, aunque con una pérdida de la convergencia en extremo rápida que puede producir el método de Newton.

La sección 10.4 describe el método de descenso más rápido. Sólo es linealmente convergente, pero no requiere las aproximaciones iniciales precisas necesarias para las técnicas más rápidas de convergencia. A menudo se usa para encontrar una buena aproximación inicial para el método de Newton o una de sus modificaciones.

En la sección 10.5 proporcionamos una introducción para los métodos de continuación, que usan un parámetro para ir de un problema con una solución fácilmente determinada a la solución del problema no lineal original.

En este capítulo se omiten muchas de las pruebas de los resultados teóricos porque implican métodos que, normalmente, se estudian en cálculo avanzado. Una buena referencia general para este material es el libro de Ortega titulado *Numerical Analysis–A Second Course (Análisis numérico – Un segundo curso)* [Or2]. Una referencia más completa es [OR].

10.1 Puntos fijos para funciones de varias variables

Un sistema de ecuaciones no lineales tiene la forma

$$\begin{aligned} f_1(x_1, x_2, \ldots, x_n) &= 0, \\ f_2(x_1, x_2, \ldots, x_n) &= 0, \\ \vdots \qquad & \quad \vdots \\ f_n(x_1, x_2, \ldots, x_n) &= 0, \end{aligned} \tag{10.1}$$

donde cada función f_i se puede pensar como un mapeo de un vector $\mathbf{x} = (x_1, x_2, \ldots, x_n)^t$ del espacio n dimensional $\mathbb{R}^n$ en la recta real $\mathbb{R}$. En la figura 10.1 se muestra una representación geométrica de un sistema no lineal cuando $n = 2$.

Este sistema de n ecuaciones no lineales en n variables también se puede representar al definir una función $\mathbf{F}$ de mapeo $\mathbb{R}^n$ en $\mathbb{R}^n$.

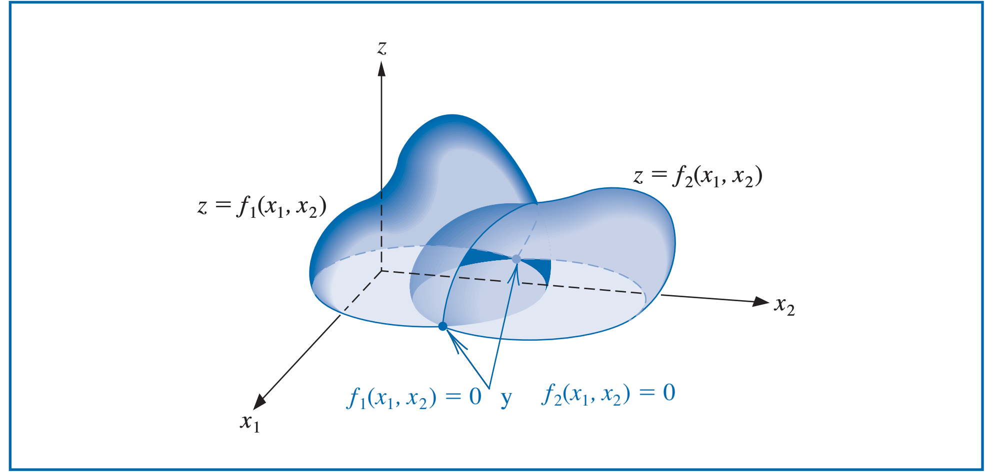
$$\mathbf{F}(x_1, x_2, \ldots, x_n) = (f_1(x_1, x_2, \ldots, x_n), f_2(x_1, x_2, \ldots, x_n), \ldots, f_n(x_1, x_2, \ldots, x_n))^t.$$

Si se utiliza notación vectorial para representar las variables $x_1, x_2, \ldots, x_n$, entonces el sistema (10.1) asume la forma

$$\mathbf{F}(\mathbf{x}) = \mathbf{0}. \tag{10.2}$$

Las funciones $f_1, f_2, \ldots, f_n$ reciben el nombre de **funciones coordenadas** de $\mathbf{F}$.

Figura 10.1



Ejemplo 1 Escriba el sistema no lineal 3 × 3

$$3x_1 - \cos(x_2x_3) - \frac{1}{2} = 0,$$

$$x_1^2 - 81(x_2 + 0.1)^2 + \operatorname{sen} x_3 + 1.06 = 0,$$

$$e^{-x_1x_2} + 20x_3 + \frac{10\pi - 3}{3} = 0,$$

en la forma (10.2)

Solución Defina las tres funciones coordenadas f_1, f_2 y f_3 desde $\mathbb{R}^3$ hasta $\mathbb{R}$ como

$$f_1(x_1, x_2, x_3) = 3x_1 - \cos(x_2x_3) - \frac{1}{2},$$

$$f_2(x_1, x_2, x_3) = x_1^2 - 81(x_2 + 0.1)^2 + \operatorname{sen} x_3 + 1.06 \quad \text{y}$$

$$f_3(x_1, x_2, x_3) = e^{-x_1x_2} + 20x_3 + \frac{10\pi - 3}{3}.$$

Entonces, defina $\mathbf{F}$ desde $\mathbb{R}^3 \to \mathbb{R}^3$ mediante

$$\begin{aligned}\mathbf{F}(\mathbf{x}) &= \mathbf{F}(x_1, x_2, x_3)\\ &= (f_1(x_1, x_2, x_3),\ f_2(x_1, x_2, x_3),\ f_3(x_1, x_2, x_3))^t\\ &= \left(3x_1 - \cos(x_2x_3) - \frac{1}{2}, x_1^2 - 81(x_2 + 0.1)^2\right.\\ &\qquad \left. + \operatorname{sen} x_3 + 1.06, e^{-x_1x_2} + 20x_3 + \frac{10\pi - 3}{3}\right)^t.\end{aligned}$$

■

Antes de analizar la solución de un sistema provisto en la forma (10.1) o (10.2), necesitamos algunos resultados respecto a la continuidad y diferenciabilidad de las funciones desde $\mathbb{R}^n$ hasta $\mathbb{R}^n$. A pesar de que este estudio podría representarse directamente (consulte el ejercicio 14), usamos un método alternativo que nos permite representar teóricamente los conceptos más difíciles de límites y continuidad en términos de funciones desde $\mathbb{R}^n$ hasta $\mathbb{R}$.

Definición 10.1 Sea f una función definida en un conjunto $D \subset \mathbb{R}^n$ y rango en $\mathbb{R}$. Se dice que la función f tiene **límite** L en $\mathbf{x}_0$, escrito

$$\lim_{\mathbf{x}\to\mathbf{x}_0} f(\mathbf{x}) = L,$$

si, dado cualquier número $\varepsilon > 0$, existe un número $\delta > 0$ con

$$|f(\mathbf{x}) - L| < \varepsilon,$$

siempre que $\mathbf{x} \in D$, y

$$0 < ||\mathbf{x} - \mathbf{x}_0|| < \delta. \quad ■$$

La existencia de un límite también es independiente de la norma vectorial particular que se usa, como se analizó en la sección 7.1. Cualquier norma conveniente se puede usar para satisfacer la condición en esta definición. El valor específico de δ dependerá de la norma seleccionada, pero la existencia de δ es independiente de la norma.

La noción de límite nos permite definir la continuidad para las funciones desde $\mathbb{R}^n$ hasta $\mathbb{R}$. Aunque es posible usar varias normas, la continuidad es independiente de la selección particular.

Definición 10.2 Sea f una función del conjunto $D \subset \mathbb{R}^n$ en $\mathbb{R}$. La función f es **continua** en $\mathbf{x}_0 \in D$ siempre que exista $\lim_{\mathbf{x}\to\mathbf{x}_0} f(\mathbf{x})$ y

$$\lim_{\mathbf{x}\to\mathbf{x}_0} f(\mathbf{x}) = f(\mathbf{x}_0).$$

Además, f es **continua** en un conjunto D si f es continua en cada punto de D. Este concepto se expresa al escribir $f \in C(D)$. ■

Las definiciones de continuidad para las funciones de n variables siguen de aquellas para una sola variable al reemplazar, siempre que sea necesario, los valores absolutos por normas.

Ahora podemos definir los conceptos de límite y continuidad para las funciones desde $\mathbb{R}^n$ hasta $\mathbb{R}^n$ al considerar las funciones coordenadas desde $\mathbb{R}^n$ en $\mathbb{R}$.

Definición 10.3 Sea $\mathbf{F}$ una función desde $D \subset \mathbb{R}^n$ a $\mathbb{R}^n$ de la forma

$$\mathbf{F}(\mathbf{x}) = (f_1(\mathbf{x}), f_2(\mathbf{x}), \ldots, f_n(\mathbf{x}))^t,$$

donde f_i es un mapeo de $\mathbb{R}^n$ hasta $\mathbb{R}$ para cada i. Definimos

$$\lim_{\mathbf{x}\to\mathbf{x}_0} \mathbf{F}(\mathbf{x}) = \mathbf{L} = (L_1, L_2, \ldots, L_n)^t,$$

si y sólo si $\lim_{\mathbf{x}\to\mathbf{x}_0} f_i(\mathbf{x}) = L_i$ para cada $i = 1, 2, \ldots, n$. ■

La función $\mathbf{F}$ es **continua** en $\mathbf{x}_0 \in D$ siempre que exista $\lim_{\mathbf{x}\to\mathbf{x}_0} \mathbf{F}(\mathbf{x})$ y $\lim_{\mathbf{x}\to\mathbf{x}_0} \mathbf{F}(\mathbf{x}) = \mathbf{F}(\mathbf{x}_0)$. Además, $\mathbf{F}$ es continua en el conjunto D si $\mathbf{F}$ es continua en cada $\mathbf{x}$ en D. Este concepto se expresa al escribir $\mathbf{F} \in C(D)$.

Para las funciones desde $\mathbb{R}$ hasta $\mathbb{R}$, la continuidad a menudo se puede evidenciar al demostrar que la función es diferenciable (consulte el teorema 1.6). Aunque este teorema se generaliza en funciones de diversas variables, la derivada (o derivada total) de una función de diversas variables está muy involucrada y no se presentará aquí. Por el contrario, establecemos el siguiente teorema, que relaciona la continuidad de una función de n variables en un punto con las derivadas parciales de la función en el punto.

Teorema 10.4 Sea f una función de $D \subset \mathbb{R}^n$ a $\mathbb{R}$ y $\mathbf{x}_0 \in D$. Suponga que existen todas las derivadas parciales de f y las constantes $\delta > 0$ y $K > 0$, de tal forma que siempre que $\|\mathbf{x} - \mathbf{x}_0\| < \delta$ y $\mathbf{x} \in D$, tenemos

$$\left| \frac{\partial f(\mathbf{x})}{\partial x_j} \right| \leq K, \quad \text{para cada } j = 1, 2, \ldots, n.$$

Entonces f es continua en $\mathbf{x}_0$. ■

Puntos fijos en $\mathbb{R}^n$

En el capítulo 2 se desarrolló un proceso iterativo para resolver una ecuación $f(x) = 0$, al transformar primero la ecuación en la forma de punto fijo $x = g(x)$. Se investigará un proceso similar para las funciones desde $\mathbb{R}^n$ hasta $\mathbb{R}^n$.

Definición 10.5 Una función $\mathbf{G}$ desde $D \subset \mathbb{R}^n$ hasta $\mathbb{R}^n$ tiene un punto fijo en $\mathbf{p} \in D$ si $\mathbf{G}(\mathbf{p}) = \mathbf{p}$. ■

El siguiente teorema generaliza el teorema de punto fijo 2.4 en la página 47 para el caso n-dimensional. Este teorema es un caso especial del teorema de función contractiva y su demostración se puede encontrar en [Or2], p. 153.

Teorema 10.6 Sea $D = \{ (x_1, x_2, \ldots, x_n)^t \mid a_i \leq x_i \leq b_i, \text{ para cada } i = 1, 2, \ldots, n \}$ para algún conjunto de constantes $a_1, a_2, \ldots, a_n$ y $b_1, b_2, \ldots, b_n$. Suponga que $\mathbf{G}$ es una función continua en $D \subset \mathbb{R}^n$ a $\mathbb{R}^n$ con la propiedad de que $\mathbf{G}(\mathbf{x}) \in D$, siempre que $\mathbf{x} \in D$. Entonces $\mathbf{G}$ tiene un punto fijo en D.

Además, suponga que todas las funciones componentes de $\mathbf{G}$ tienen derivadas parciales continuas y que existe una constante $K < 1$ con

$$\left| \frac{\partial g_i(\mathbf{x})}{\partial x_j} \right| \leq \frac{K}{n}, \quad \text{siempre que} \quad \mathbf{x} \in D,$$

para cada $j = 1, 2, \ldots, n$ y cada función componente g_i. Entonces, la sucesión de punto fijo $\{\mathbf{x}^{(k)}\}_{k=0}^{\infty}$ definida por $\mathbf{x}^{(0)}$ seleccionada arbitrariamente en D y generada por medio de

$$\mathbf{x}^{(k)} = G(\mathbf{x}^{(k-1)}), \quad \text{para cada } k \geq 1$$

converge al único punto fijo $\mathbf{p} \in D$ y

$$\left\|\mathbf{x}^{(k)} - \mathbf{p}\right\|_\infty \leq \frac{K^k}{1-K} \left\|\mathbf{x}^{(1)} - \mathbf{x}^{(0)}\right\|_\infty . \tag{10.3}$$

■

Ejemplo 2 Considere el sistema no lineal

$$3x_1 - \cos(x_2 x_3) - \frac{1}{2} = 0,$$
$$x_1^2 - 81(x_2 + 0.1)^2 + \operatorname{sen} x_3 + 1.06 = 0, \text{ y}$$
$$e^{-x_1 x_2} + 20x_3 + \frac{10\pi - 3}{3} = 0$$

en forma de punto fijo $\mathbf{x} = \mathbf{G}(\mathbf{x})$ al resolver la i-ésima ecuación para x_i. Muestre que existe una solución única en

$$D = \{ (x_1, x_2, x_3)^t \mid -1 \leq x_i \leq 1, \quad \text{para cada } i = 1, 2, 3\},$$

e itere a partir de $\mathbf{x}^{(0)} = (0.1, 0.1, -0.1)^t$ hasta una precisión dentro de 10^{-5} en la norma l_∞ obtenida.

Solución Al resolver la i-ésima ecuación para x_i obtenemos el problema de punto fijo

$$
\begin{aligned}
x_1 &= \frac{1}{3}\cos(x_2x_3) + \frac{1}{6},\\
x_2 &= \frac{1}{9}\sqrt{x_1^2 + \operatorname{sen} x_3 + 1.06} - 0.1, \qquad (10.4)\\
x_3 &= -\frac{1}{20}e^{-x_1x_2} - \frac{10\pi - 3}{60}.
\end{aligned}
$$

Sea $\mathbf{G} : \mathbb{R}^3 \to \mathbb{R}^3$ definido por $\mathbf{G}(\mathbf{x}) = (g_1(\mathbf{x}), g_2(\mathbf{x}), g_3(\mathbf{x}))^t$, donde

$$
\begin{aligned}
g_1(x_1, x_2, x_3) &= \frac{1}{3}\cos(x_2x_3) + \frac{1}{6},\\
g_2(x_1, x_2, x_3) &= \frac{1}{9}\sqrt{x_1^2 + \operatorname{sen} x_3 + 1.06} - 0.1,\\
g_3(x_1, x_2, x_3) &= -\frac{1}{20}e^{-x_1x_2} - \frac{10\pi - 3}{60}.
\end{aligned}
$$

Los teoremas 10.4 y 10.6 se usarán para mostrar que $\mathbf{G}$ tiene un punto fijo único en

$$D = \{\, (x_1, x_2, x_3)^t \mid -1 \le x_i \le 1, \quad \text{para cada } i = 1, 2, 3\}.$$

Para $\mathbf{x} = (x_1, x_2, x_3)^t$ en D,

$$
\begin{aligned}
|g_1(x_1, x_2, x_3)| &\le \frac{1}{3}|\cos(x_2x_3)| + \frac{1}{6} \le 0.50,\\
|g_2(x_1, x_2, x_3)| &= \left|\frac{1}{9}\sqrt{x_1^2 + \operatorname{sen} x_3 + 1.06} - 0.1\right| \le \frac{1}{9}\sqrt{1 + \operatorname{sen} 1 + 1.06} - 0.1 < 0.09,
\end{aligned}
$$

y

$$|g_3(x_1, x_2, x_3)| = \frac{1}{20}e^{-x_1x_2} + \frac{10\pi - 3}{60} \le \frac{1}{20}e + \frac{10\pi - 3}{60} < 0.61.$$

Por lo que tenemos, para cada $i = 1, 2, 3$,

$$-1 \le g_i(x_1, x_2, x_3) \le 1.$$

Por lo tanto, $\mathbf{G}(\mathbf{x}) \in D$ siempre que $\mathbf{x} \in D$.

Encontrar límites para las derivadas parciales en D nos da

$$\left|\frac{\partial g_1}{\partial x_1}\right| = 0, \quad \left|\frac{\partial g_2}{\partial x_2}\right| = 0, \quad \text{y} \quad \left|\frac{\partial g_3}{\partial x_3}\right| = 0$$

así como

$$
\begin{aligned}
&\left|\frac{\partial g_1}{\partial x_2}\right| \le \frac{1}{3}|x_3| \cdot |\operatorname{sen} x_2x_3| \le \frac{1}{3}\operatorname{sen} 1 < 0.281, \quad \left|\frac{\partial g_1}{\partial x_3}\right| \le \frac{1}{3}|x_2| \cdot |\operatorname{sen} x_2x_3| \le \frac{1}{3}\operatorname{sen} 1 < 0.281,\\
&\left|\frac{\partial g_2}{\partial x_1}\right| = \frac{|x_1|}{9\sqrt{x_1^2 + \operatorname{sen} x_3 + 1.06}} < \frac{1}{9\sqrt{0.218}} < 0.238,\\
&\left|\frac{\partial g_2}{\partial x_3}\right| = \frac{|\cos x_3|}{18\sqrt{x_1^2 + \operatorname{sen} x_3 + 1.06}} < \frac{1}{18\sqrt{0.218}} < 0.119,\\
&\left|\frac{\partial g_3}{\partial x_1}\right| = \frac{|x_2|}{20}e^{-x_1x_2} \le \frac{1}{20}e < 0.14, \quad \text{y} \quad \left|\frac{\partial g_3}{\partial x_2}\right| = \frac{|x_1|}{20}e^{-x_1x_2} \le \frac{1}{20}e < 0.14.
\end{aligned}
$$

Todas las derivadas parciales de g_1, g_2 y g_3 están acotadas en D, por lo que el teorema 10.4 implica que estas funciones son continuas en D. Por consiguiente, **G** es continua en D. Además, para cada $\mathbf{x} \in D$,

$$\left|\frac{\partial g_i(\mathbf{x})}{\partial x_j}\right| \leq 0.281, \quad \text{para cada } i = 1, 2, 3 \quad \text{y} \quad j = 1, 2, 3,$$

y la condición en la segunda parte del teorema 10.6 se mantiene con $K = 3(0.281) = 0.843$.

De la misma forma, también es posible mostrar que $\partial g_i/\partial x_j$ es continua en D para cada $i = 1, 2, 3$ y $j = 1, 2, 3$. (Esto se considera en el ejercicio 13.) Por consiguiente, **G** tiene un único punto fijo en D y el sistema no lineal tiene una solución en D.

Observe que **G** tiene un punto fijo único en D, y esto no implica que la solución para el sistema original sea la única en este dominio porque la solución para x_2 en la ecuación (10.4) implicaba la selección de la raíz cuadrada principal. El ejercicio 5d) examina lo que ocurre si, por el contrario, se selecciona la raíz cuadrada negativa en este paso.

Para aproximar el punto fijo **p**, seleccionamos $\mathbf{x}^{(0)} = (0.1, 0.1, -0.1)^t$. La sucesión de vectores generados por

$$x_1^{(k)} = \frac{1}{3}\cos x_2^{(k-1)} x_3^{(k-1)} + \frac{1}{6},$$

$$x_2^{(k)} = \frac{1}{9}\sqrt{\left(x_1^{(k-1)}\right)^2 + \operatorname{sen} x_3^{(k-1)} + 1.06} - 0.1, \text{ y}$$

$$x_3^{(k)} = -\frac{1}{20} e^{-x_1^{(k-1)} x_2^{(k-1)}} - \frac{10\pi - 3}{60}$$

converge con la solución única del sistema en la ecuación (10.4) Los resultados en la tabla 10.1 se generaron hasta que

$$\left\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\right\|_\infty < 10^{-5}.$$

■

Tabla 10.1

k	$x_1^{(k)}$	$x_2^{(k)}$	$x_3^{(k)}$	$\left\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\right\|_\infty$
0	0.10000000	0.10000000	−0.10000000	
1	0.49998333	0.00944115	−0.52310127	0.423
2	0.49999593	0.00002557	−0.52336331	9.4×10^{-3}
3	0.50000000	0.00001234	−0.52359814	2.3×10^{-4}
4	0.50000000	0.00000003	−0.52359847	1.2×10^{-5}
5	0.50000000	0.00000002	−0.52359877	3.1×10^{-7}

Podríamos usar la cota de error (10.3) con $K = 0.843$ en el ejemplo previo. Esto da

$$\|\mathbf{x}^{(5)} - \mathbf{p}\|_\infty \leq \frac{(0.843)^5}{1 - 0.843}(0.423) < 1.15,$$

lo que no indica la verdadera precisión de $\mathbf{x}^{(5)}$. La solución real es

$$\mathbf{p} = \left(0.5, 0, -\frac{\pi}{6}\right)^t \approx (0.5, 0, -0.5235987757)^t, \quad \text{por lo que} \quad \|\mathbf{x}^{(5)} - \mathbf{p}\|_\infty \leq 2 \times 10^{-8}.$$

Aceleración de la convergencia

Una forma de acelerar la convergencia de la iteración de punto fijo es usar los últimos cálculos $x_1^{(k)}, \ldots, x_{i-1}^{(k)}$ en lugar de $x_1^{(k-1)}, \ldots, x_{i-1}^{(k-1)}$ para calcular $x_i^{(k)}$, como en el método de Gauss-Siedel para los sistemas lineales. Entonces, las ecuaciones del componente para el problema en el ejemplo se convierten en

$$x_1^{(k)} = \frac{1}{3}\cos\left(x_2^{(k-1)}x_3^{(k-1)}\right) + \frac{1}{6},$$

$$x_2^{(k)} = \frac{1}{9}\sqrt{\left(x_1^{(k)}\right)^2 + \operatorname{sen} x_3^{(k-1)} + 1.06} - 0.1, \text{ y}$$

$$x_3^{(k)} = -\frac{1}{20}e^{-x_1^{(k)}x_2^{(k)}} - \frac{10\pi - 3}{60}.$$

Con $\mathbf{x}^{(0)} = (0.1, 0.1, -0.1)^t$, los resultados de estos cálculos se muestran en la tabla 10.2

Tabla 10.2

k	$x_1^{(k)}$	$x_2^{(k)}$	$x_3^{(k)}$	$\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\|_\infty$
0	0.10000000	0.10000000	−0.10000000	
1	0.49998333	0.02222979	−0.52304613	0.423
2	0.49997747	0.00002815	−0.52359807	2.2×10^{-2}
3	0.50000000	0.00000004	−0.52359877	2.8×10^{-5}
4	0.50000000	0.00000000	−0.52359877	3.8×10^{-8}

La iteración $\mathbf{x}^{(4)}$ es precisa dentro de 10^{-7} en la norma l_∞; por lo que la convergencia estaba, de hecho, acelerada para este problema al usar el método de Gauss-Siedel. Sin embargo, este método no *siempre* acelera la convergencia.

La sección Conjunto de ejercicios 10.1 está disponible en línea. Encuentre la ruta de acceso en las páginas preliminares.

10.2 Método de Newton

El problema en el ejemplo 2 de la sección 10.1 se transforma en un problema de punto fijo convergente al resolver algebraicamente las tres ecuaciones para las tres variables x_1, x_2 y x_3. Sin embargo, es poco común ser capaz de encontrar una representación explícita para todas las variables. En esta sección, consideramos un procedimiento algorítmico para realizar la transformación en una situación más general.

Para construir el algoritmo que conduce a un método de punto fijo adecuado en el caso unidimensional, encontramos una función ϕ con la propiedad de que

$$g(x) = x - \phi(x)f(x)$$

da convergencia cuadrática para el punto fijo p de la función g (consulte la sección 2.4). A partir de esta condición el método de Newton evolucionó al seleccionar $\phi(x) = 1/f'(x)$, suponiendo que $f'(x) \neq 0$.

Un enfoque similar en el caso n-dimensional implica una matriz

$$A(\mathbf{x}) = \begin{bmatrix} a_{11}(\mathbf{x}) & a_{12}(\mathbf{x}) & \cdots & a_{1n}(\mathbf{x}) \\ a_{21}(\mathbf{x}) & a_{22}(\mathbf{x}) & \cdots & a_{2n}(\mathbf{x}) \\ \vdots & \vdots & & \vdots \\ a_{n1}(\mathbf{x}) & a_{n2}(\mathbf{x}) & \cdots & a_{nn}(\mathbf{x}) \end{bmatrix}, \tag{10.5}$$

donde cada una de las entradas $a_{ij}(\mathbf{x})$ es una función de $\mathbb{R}^n$ a $\mathbb{R}$. Esto requiere encontrar $A(\mathbf{x})$ de tal forma que

$$\mathbf{G}(\mathbf{x}) = \mathbf{x} - A(\mathbf{x})^{-1}\mathbf{F}(\mathbf{x})$$

da convergencia cuadrática para la solución de $\mathbf{F}(\mathbf{x}) = \mathbf{0}$, suponiendo que $A(\mathbf{x})$ es no singular en el punto fijo $\mathbf{p}$ de $\mathbf{G}$.

El siguiente teorema se compara con el teorema 2.8 en la página 59. Su demostración requiere ser capaz de expresar $\mathbf{G}$ en términos de su serie de Taylor en n variables alrededor de $\mathbf{p}$.

Teorema 10.7 Si $\mathbf{p}$ es la solución de $\mathbf{G}(\mathbf{x}) = \mathbf{x}$. Suponga que existe un número $\delta > 0$ con las propiedades:

i) $\partial g_i/\partial x_j$ es continua en $N_\delta = \{\mathbf{x} \mid \|\mathbf{x} - \mathbf{p}\| < \delta\}$, para cada $i = 1, 2, \ldots, n$ y $j = 1, 2, \ldots, n$;

ii) $\partial^2 g_i(\mathbf{x})/(\partial x_j \partial x_k)$ es continua y $|\partial^2 g_i(\mathbf{x})/(\partial x_j \partial x_k)| \leq M$ para algunas constantes M, siempre que $\mathbf{x} \in N_\delta$, para cada $i = 1, 2, ..., n$, $j = 1, 2, \ldots, n$, y $k = 1, 2, \ldots, n$;

iii) $\partial g_i(\mathbf{p})/\partial x_k = 0$, para cada $i = 1, 2, \ldots, n$ y $k = 1, 2, \ldots, n$.

Entonces, un número $\hat{\delta} \leq \delta$ existe de tal forma que la sucesión generada por $\mathbf{x}^{(k)} = \mathbf{G}(\mathbf{x}^{(k-1)})$ converge de forma cuadrática en $\mathbf{p}$ para cualquier selección de $\mathbf{x}^{(0)}$, siempre y cuando $\left\|\mathbf{x}^{(0)} - \mathbf{p}\right\| < \hat{\delta}$. Además,

$$\|\mathbf{x}^{(k)} - \mathbf{p}\|_\infty \leq \frac{n^2 M}{2}\|\mathbf{x}^{(k-1)} - \mathbf{p}\|_\infty^2, \quad \text{para cada } k \geq 1. \quad ■$$

Para aplicar el teorema 10.7, suponga que $A(\mathbf{x})$ es una matriz $n \times n$ de funciones de $\mathbb{R}^n$ a $\mathbb{R}$ en la forma de la ecuación (10.5), donde las entradas específicas se seleccionarán más adelante. Suponga, además, que $A(\mathbf{x})$ es no singular cerca de una solución $\mathbf{p}$ de $\mathbf{F}(\mathbf{x}) = \mathbf{0}$ y sea que $b_{ij}(\mathbf{x})$ denote la entrada de $A(\mathbf{x})^{-1}$ en la i-ésima fila y la j-ésima columna.

Para $\mathbf{G}(\mathbf{x}) = \mathbf{x} - A(\mathbf{x})^{-1}\mathbf{F}(\mathbf{x})$, tenemos $g_i(\mathbf{x}) = x_i - \sum_{j=1}^{n} b_{ij}(\mathbf{x}) f_j(\mathbf{x})$. De modo que,

$$\frac{\partial g_i}{\partial x_k}(\mathbf{x}) = \begin{cases} 1 - \displaystyle\sum_{j=1}^{n}\left(b_{ij}(\mathbf{x})\frac{\partial f_j}{\partial x_k}(\mathbf{x}) + \frac{\partial b_{ij}}{\partial x_k}(\mathbf{x}) f_j(\mathbf{x})\right), & \text{si } i = k, \\ -\displaystyle\sum_{j=1}^{n}\left(b_{ij}(\mathbf{x})\frac{\partial f_j}{\partial x_k}(\mathbf{x}) + \frac{\partial b_{ij}}{\partial x_k}(\mathbf{x}) f_j(\mathbf{x})\right), & \text{si } i \neq k. \end{cases}$$

El teorema 10.7 implica que necesitamos $\partial g_i(\mathbf{p})/\partial x_k = 0$, para cada $i = 1, 2, \ldots, n$ y $k = 1, 2, \ldots, n$. Esto significa que para $i = k$,

$$0 = 1 - \sum_{j=1}^{n} b_{ij}(\mathbf{p})\frac{\partial f_j}{\partial x_i}(\mathbf{p}),$$

es decir,

$$\sum_{j=1}^{n} b_{ij}(\mathbf{p})\frac{\partial f_j}{\partial x_i}(\mathbf{p}) = 1. \tag{10.6}$$

Cuando $k \neq i$,

$$0 = -\sum_{j=1}^{n} b_{ij}(\mathbf{p})\frac{\partial f_j}{\partial x_k}(\mathbf{p}),$$

de tal forma que

$$\sum_{j=1}^{n} b_{ij}(\mathbf{p}) \frac{\partial f_j}{\partial x_k}(\mathbf{p}) = 0. \tag{10.7}$$

La matriz jacobiana

Defina la matriz $J(\mathbf{x})$ mediante

$$J(\mathbf{x}) = \begin{bmatrix} \frac{\partial f_1}{\partial x_1}(\mathbf{x}) & \frac{\partial f_1}{\partial x_2}(\mathbf{x}) & \cdots & \frac{\partial f_1}{\partial x_n}(\mathbf{x}) \\ \frac{\partial f_2}{\partial x_1}(\mathbf{x}) & \frac{\partial f_2}{\partial x_2}(\mathbf{x}) & \cdots & \frac{\partial f_2}{\partial x_n}(\mathbf{x}) \\ \vdots & \vdots & & \vdots \\ \frac{\partial f_n}{\partial x_1}(\mathbf{x}) & \frac{\partial f_n}{\partial x_2}(\mathbf{x}) & \cdots & \frac{\partial f_n}{\partial x_n}(\mathbf{x}) \end{bmatrix}. \tag{10.8}$$

Entonces, las condiciones (10.6) y (10.7) requieren que

$$A(\mathbf{p})^{-1} J(\mathbf{p}) = I, \text{ la matriz de identidad,} \quad \text{por lo que} \quad A(\mathbf{p}) = J(\mathbf{p}).$$

Una selección adecuada para $A(\mathbf{x})$ es, por consiguiente, $A(\mathbf{x}) = J(\mathbf{x})$ ya que ésta satisface la condición iii) en el teorema 10.7. La función $\mathbf{G}$ se define mediante

$$\mathbf{G}(\mathbf{x}) = \mathbf{x} - J(\mathbf{x})^{-1}\mathbf{F}(\mathbf{x}),$$

y el procedimiento de iteración de punto fijo evoluciona al seleccionar $\mathbf{x}^{(0)}$ y generar, para $k \geq 1$,

$$\mathbf{x}^{(k)} = \mathbf{G}(\mathbf{x}^{(k-1)}) = \mathbf{x}^{(k-1)} - J(\mathbf{x}^{(k-1)})^{-1}\mathbf{F}(\mathbf{x}^{(k-1)}). \tag{10.9}$$

Esto recibe el nombre de **método de Newton para sistemas no lineales** y en general se espera que proporcione convergencia cuadrática, siempre y cuando se conozca un valor inicial suficientemente preciso y que $J(\mathbf{p})^{-1}$ exista. La matriz $J(\mathbf{x})$ recibe el nombre de matriz **jacobiana** y tiene numerosas aplicaciones en análisis. En especial, podría resultar familiar para el lector debido a su aplicación en la integración múltiple de una función de diversas variables sobre una región que requiere que se efectúe un cambio de variables.

La matriz jacobiana apareció por primera vez en 1815, en un artículo de Cauchy, pero Jacobi escribió *De determinantibus functionalibus*, en 1841, y probó numerosos resultados sobre esta matriz.

Una debilidad en el método de Newton surge de la necesidad de calcular e invertir la matriz $J(\mathbf{x})$ en cada paso. En la práctica, el cálculo explícito de $J(\mathbf{x})^{-1}$ se evita al realizar la operación en una forma de dos pasos. Primero, se encuentra un vector $\mathbf{y}$ que satisface $J(\mathbf{x}^{(k-1)})\mathbf{y} = -\mathbf{F}(\mathbf{x}^{(k-1)})$. Entonces, la nueva aproximación, $\mathbf{x}^{(k)}$, se obtiene sumando $\mathbf{y}$ a $\mathbf{x}^{(k-1)}$. El algoritmo 10.1 utiliza este procedimiento de dos pasos.

ALGORITMO 10.1

Método de Newton para sistemas

Para aproximar la solución del sistema no lineal $\mathbf{F}(\mathbf{x}) = \mathbf{0}$, dada una aproximación inicial $\mathbf{x}$:

ENTRADA número n de ecuaciones y valores desconocidos; aproximación inicial $\mathbf{x} = (x_1, \ldots, x_n)^t$, tolerancia TOL; número máximo de iteraciones N.

SALIDA solución aproximada $\mathbf{x} = (x_1, \ldots, x_n)^t$ o un mensaje que indica que se excedió el número de iteraciones.



Paso 1 Determine $k = 1$.

Paso 2 Mientras ($k \leq N$) haga los pasos 3–7.

Paso 3 Calcule $\mathbf{F}(\mathbf{x})$ y $J(\mathbf{x})$, donde $J(\mathbf{x})_{i,j} = (\partial f_i(\mathbf{x})/\partial x_j)$ para $1 \leq i, j \leq n$.

Paso 4 Resuelva el sistema lineal $n \times n$ $J(\mathbf{x})\mathbf{y} = -\mathbf{F}(\mathbf{x})$.

Paso 5 Determine $\mathbf{x} = \mathbf{x} + \mathbf{y}$.

Paso 6 Si $||\mathbf{y}|| < TOL$ entonces SALIDA ($\mathbf{x}$);
(*El procedimiento fue exitoso*.)
PARE.

Paso 7 Determine $k = k + 1$.

Paso 8 SALIDA ('Número máximo de iteraciones excedido');
(*El procedimiento no fue exitoso*.)
PARE. ■

Ejemplo 1 El sistema no lineal

$$3x_1 - \cos(x_2x_3) - \frac{1}{2} = 0,$$
$$x_1^2 - 81(x_2 + 0.1)^2 + \operatorname{sen} x_3 + 1.06 = 0, \text{ y}$$
$$e^{-x_1x_2} + 20x_3 + \frac{10\pi - 3}{3} = 0$$

se mostró en el ejemplo 2 de la sección 10.1 para tener la solución aproximada $(0.5, 0, -0.52359877)^t$. Aplique el método de Newton para este problema con $\mathbf{x}^{(0)} = (0.1, 0.1, -0.1)^t$.

Solución Defina

$$\mathbf{F}(x_1, x_2, x_3) = (f_1(x_1, x_2, x_3), f_2(x_1, x_2, x_3), f_3(x_1, x_2, x_3))^t,$$

donde

$$f_1(x_1, x_2, x_3) = 3x_1 - \cos(x_2x_3) - \frac{1}{2},$$
$$f_2(x_1, x_2, x_3) = x_1^2 - 81(x_2 + 0.1)^2 + \operatorname{sen} x_3 + 1.06,$$

y

$$f_3(x_1, x_2, x_3) = e^{-x_1x_2} + 20x_3 + \frac{10\pi - 3}{3}.$$

La matriz jacobiana $J(\mathbf{x})$ para este sistema es

$$J(x_1, x_2, x_3) = \begin{bmatrix} 3 & x_3 \operatorname{sen} x_2x_3 & x_2 \operatorname{sen} x_2x_3 \\ 2x_1 & -162(x_2 + 0.1) & \cos x_3 \\ -x_2e^{-x_1x_2} & -x_1e^{-x_1x_2} & 20 \end{bmatrix}.$$

Sea $\mathbf{x}^{(0)} = (0.1, 0.1, -0.1)^t$. Entonces $\mathbf{F}(\mathbf{x}^{(0)} = (-0.199995, -2.269833417, 8.462025346)^t$ y

$$J(\mathbf{x}^{(0)}) = \begin{bmatrix} 3 & 9.999833334 \times 10^{-4} & 9.999833334 \times 10^{-4} \\ 0.2 & -32.4 & 0.9950041653 \\ -0.09900498337 & -0.09900498337 & 20 \end{bmatrix}.$$

La resolución del sistema lineal $J(\mathbf{x}^{(0)})\mathbf{y}^{(0)} = -\mathbf{F}(\mathbf{x}^{(0)})$ da

$$\mathbf{y}^{(0)} = \begin{bmatrix} 0.3998696728 \\ -0.08053315147 \\ -0.4215204718 \end{bmatrix} \quad \text{y} \quad \mathbf{x}^{(1)} = \mathbf{x}^{(0)} + \mathbf{y}^{(0)} = \begin{bmatrix} 0.4998696782 \\ 0.01946684853 \\ -0.5215204718 \end{bmatrix}.$$

Al continuar para $k = 2, 3, \ldots$, tenemos

$$\begin{bmatrix} x_1^{(k)} \\ x_2^{(k)} \\ x_3^{(k)} \end{bmatrix} = \begin{bmatrix} x_1^{(k-1)} \\ x_2^{(k-1)} \\ x_3^{(k-1)} \end{bmatrix} + \begin{bmatrix} y_1^{(k-1)} \\ y_2^{(k-1)} \\ y_3^{(k-1)} \end{bmatrix},$$

donde

$$\begin{bmatrix} y_1^{(k-1)} \\ y_2^{(k-1)} \\ y_3^{(k-1)} \end{bmatrix} = -\left(J\left(x_1^{(k-1)}, x_2^{(k-1)}, x_3^{(k-1)}\right)\right)^{-1} \mathbf{F}\left(x_1^{(k-1)}, x_2^{(k-1)}, x_3^{(k-1)}\right).$$

Por lo tanto, en el k-ésimo paso, el sistema lineal $J\left(\mathbf{x}^{(k-1)}\right)\mathbf{y}^{(k-1)} = -\mathbf{F}\left(\mathbf{x}^{(k-1)}\right)$ se debe resolver, donde

$$J\left(\mathbf{x}^{(k-1)}\right) = \begin{bmatrix} 3 & x_3^{(k-1)} \operatorname{sen} x_2^{(k-1)} x_3^{(k-1)} & x_2^{(k-1)} \operatorname{sen} x_2^{(k-1)} x_3^{(k-1)} \\ 2x_1^{(k-1)} & -162\left(x_2^{(k-1)} + 0.1\right) & \cos x_3^{(k-1)} \\ -x_2^{(k-1)} e^{-x_1^{(k-1)} x_2^{(k-1)}} & -x_1^{(k-1)} e^{-x_1^{(k-1)} x_2^{(k-1)}} & 20 \end{bmatrix},$$

$$\mathbf{y}^{(k-1)} = \begin{bmatrix} y_1^{(k-1)} \\ y_2^{(k-1)} \\ y_3^{(k-1)} \end{bmatrix},$$

y

$$\mathbf{F}\left(\mathbf{x}^{(k-1)}\right) = \begin{bmatrix} 3x_1^{(k-1)} - \cos x_2^{(k-1)} x_3^{(k-1)} - \frac{1}{2} \\ \left(x_1^{(k-1)}\right)^2 - 81\left(x_2^{(k-1)} + 0.1\right)^2 + \operatorname{sen} x_3^{(k-1)} + 1.06 \\ e^{-x_1^{(k-1)} x_2^{(k-1)}} + 20x_3^{(k-1)} + \frac{10\pi - 3}{3} \end{bmatrix}.$$

Los resultados por medio del procedimiento iterativo se muestran en la tabla 10.3. ■

Table 10.3

k	$x_1^{(k)}$	$x_2^{(k)}$	$x_3^{(k)}$	$\lVert \mathbf{x}^{(k)} - \mathbf{x}^{(k-1)} \rVert_\infty$
0	0.1000000000	0.1000000000	−0.1000000000	
1	0.4998696728	0.0194668485	−0.5215204718	0.4215204718
2	0.5000142403	0.0015885914	−0.5235569638	1.788×10^{-2}
3	0.5000000113	0.0000124448	−0.5235984500	1.576×10^{-3}
4	0.5000000000	8.516×10^{-10}	−0.5235987755	1.244×10^{-5}
5	0.5000000000	-1.375×10^{-11}	−0.5235987756	8.654×10^{-10}

El ejemplo previo ilustra que el método de Newton puede converger muy rápidamente una vez que se obtiene una buena aproximación que está cerca de la solución verdadera. Sin embargo, no siempre es fácil determinar buenos valores iniciales, y el uso del método es comparativamente caro. En la siguiente sección consideramos un método para superar

la última debilidad. Normalmente se encuentran buenos valores iniciales con el método de descenso más rápido, que se analizará en la sección 10.4.

La sección Conjunto de ejercicios 10.2 está disponible en línea. Encuentre la ruta de acceso en las páginas preliminares.

10.3 Métodos cuasi-Newton

Una debilidad significativa del método de Newton para resolver sistemas de ecuaciones no lineales es la necesidad, en cada iteración, de determinar una matriz y resolver un sistema lineal $n \times n$ que implica esta matriz. Considere la cantidad de cálculos relacionados con una iteración del método de Newton. La matriz jacobiana relacionada con un sistema de n ecuaciones lineales escrito en la forma $\mathbf{F}(\mathbf{x}) = \mathbf{0}$ requiere que las derivadas parciales n^2 de las n funciones de componentes de $\mathbf{F}$ sean determinadas y evaluadas. En muchas situaciones, la evaluación exacta de las derivadas parciales es inconveniente aunque el problema se ha vuelto más tratable con el uso generalizado de los sistemas computacionales simbólicos, como Maple, Mathematica y Matlab.

Cuando la evaluación exacta no es práctica, podemos usar aproximaciones de diferencia finita para las derivadas parciales. Por ejemplo,

$$\frac{\partial f_j}{\partial x_k}(\mathbf{x}^{(i)}) \approx \frac{f_j(\mathbf{x}^{(i)} + \mathbf{e}_k h) - f_j(\mathbf{x}^{(i)})}{h}, \tag{10.10}$$

donde h es pequeño en valor absoluto y $\mathbf{e}_k$ es el vector cuya única entrada diferente a cero es un 1 en la k-ésima coordenada. Sin embargo, esta aproximación sigue requiriendo la realización de por lo menos n^2 evaluaciones funcionales escalares para aproximar la matriz jacobiana y no disminuye la cantidad de cálculos, en general $O(n^3)$, requerida para resolver el sistema lineal relacionado con esta matriz jacobiana aproximada.

El esfuerzo computacional total para una sola iteración del método de Newton es, por consiguiente, por lo menos $n^2 + n$ evaluaciones funcionales escalares (n^2 para la evaluación de la matriz jacobiana y n para la evaluación de $\mathbf{F}$) junto con $O(n^3)$ operaciones aritméticas para resolver el sistema lineal. Esta cantidad de esfuerzo computacional es amplia, excepto para los valores relativamente pequeños de n y funciones escalares fácilmente evaluadas.

En esta sección consideramos una generalización del método de secante para los sistemas de ecuaciones no lineales, una técnica conocida como **método de Broyden** (consulte [Broy]). El método sólo requiere n evaluaciones funcionales escalares por iteración y también reduce el número de cálculos aritméticos para $O(n^2)$. Pertenece a una clase de métodos conocidos como *actualizaciones secantes del cambio mínimo* que produce algoritmos llamados **cuasi-Newton**. Estos métodos reemplazan la matriz jacobiana en el método de Newton con una matriz de aproximación que se actualiza fácilmente en cada iteración.

La desventaja de los métodos cuasi-Newton es que la convergencia cuadrática del método de Newton se pierde, al ser reemplazada, en general, mediante una convergencia llamada **superlineal**. Esto implica que

$$\lim_{i \to \infty} \frac{\left\|\mathbf{x}^{(i+1)} - \mathbf{p}\right\|}{\left\|\mathbf{x}^{(i)} - \mathbf{p}\right\|} = 0,$$

donde $\mathbf{p}$ denota la solución para $\mathbf{F}(\mathbf{x}) = \mathbf{0}$ y $\mathbf{x}^{(i)}$ y $\mathbf{x}^{(i+1)}$ son aproximaciones consecutivas para $\mathbf{p}$.

En muchas aplicaciones, la reducción de la convergencia superlineal es un intercambio más que aceptable para el decremento de la cantidad de cálculos. Una desventaja adicional de los métodos cuasi-Newton es que, a diferencia del método de Newton, no se autocorrigen.

En general, el método de Newton, corregirá el error de redondeo con iteraciones sucesivas, pero a menos que se incluyan resguardos especiales, el método de Broyden no lo hará.

Para describir el método de Broyden, suponga que se determina una aproximación inicial $\mathbf{x}^{(0)}$ para la solución $\mathbf{p}$ de $\mathbf{F}(\mathbf{x}) = \mathbf{0}$. Calculamos la siguiente aproximación $\mathbf{x}^{(1)}$ de la misma forma que el método de Newton. Si es inconveniente determinar $J(\mathbf{x}^{(0)})$ exactamente, usamos las ecuaciones de diferencia dadas por la ecuación (10.10) para aproximar las derivadas parciales. Para calcular $\mathbf{x}^{(2)}$, sin embargo, partimos del método de Newton y examinamos el método de secante para una ecuación no lineal singular. El método de secante usa la aproximación

$$f'(x_1) \approx \frac{f(x_1) - f(x_0)}{x_1 - x_0}$$

como reemplazo para $f'(x_1)$ en el método de Newton de una sola variable.

Para los sistemas no lineales, $\mathbf{x}^{(1)} - \mathbf{x}^{(0)}$ es un vector, por lo que el cociente correspondiente es indefinido. Sin embargo, el método procede de modo parecido en que reemplazamos la matriz $J\left(\mathbf{x}^{(1)}\right)$ en el método de Newton para sistemas por medio de una matriz A_1 con la propiedad de que

$$A_1\left(\mathbf{x}^{(1)} - \mathbf{x}^{(0)}\right) = \mathbf{F}\left(\mathbf{x}^{(1)}\right) - \mathbf{F}\left(\mathbf{x}^{(0)}\right). \tag{10.11}$$

Cualquier vector diferente de cero en $\mathbb{R}^n$ se puede escribir como la suma de un múltiplo de $\mathbf{x}^{(1)} - \mathbf{x}^{(0)}$ y un múltiplo de un vector en el complemento ortogonal de $\mathbf{x}^{(1)} - \mathbf{x}^{(0)}$ (consulte el ejercicio 10). Por lo que, para definir únicamente la matriz A_1, también necesitamos especificar cómo actúa en el complemento ortogonal de $\mathbf{x}^{(1)} - \mathbf{x}^{(0)}$. No existe información disponible sobre el cambio en $\mathbf{F}$ en una dirección ortogonal para $\mathbf{x}^{(1)} - \mathbf{x}^{(0)}$, por lo que especificamos que no se puede realizar ningún cambio en esta dirección; es decir,

$$A_1\mathbf{z} = J(\mathbf{x}^{(0)})\mathbf{z}, \quad \text{siempre que} \left(\mathbf{x}^{(1)} - \mathbf{x}^{(0)}\right)^t \mathbf{z} = 0. \tag{10.12}$$

Por lo tanto, cualquier vector ortogonal para $\mathbf{x}^{(1)} - \mathbf{x}^{(0)}$ no resulta afectado por la actualización de $J(\mathbf{x}^{(0)})$, que se usó para calcular $\mathbf{x}^{(1)}$ para A_1, lo cual se usa en la determinación de $\mathbf{x}^{(2)}$.

Las condiciones (10.11) y (10.12) definen únicamente A_1 (consulte [DM]) como

$$A_1 = J(\mathbf{x}^{(0)}) + \frac{\left[\mathbf{F}\left(\mathbf{x}^{(1)}\right) - \mathbf{F}\left(\mathbf{x}^{(0)}\right) - J\left(\mathbf{x}^{(0)}\right)\left(\mathbf{x}^{(1)} - \mathbf{x}^{(0)}\right)\right]\left(\mathbf{x}^{(1)} - \mathbf{x}^{(0)}\right)^t}{\left\|\mathbf{x}^{(1)} - \mathbf{x}^{(0)}\right\|_2^2}.$$

Es la matriz que se usa en lugar de $J\left(\mathbf{x}^{(1)}\right)$ para determinar $\mathbf{x}^{(2)}$ como

$$\mathbf{x}^{(2)} = \mathbf{x}^{(1)} - A_1^{-1}\mathbf{F}\left(\mathbf{x}^{(1)}\right).$$

Una vez que se ha determinado $\mathbf{x}^{(2)}$, el método se repite para hallar $\mathbf{x}^{(3)}$, por medio de A_1 en lugar de $A_0 \equiv J\left(\mathbf{x}^{(0)}\right)$ y con $\mathbf{x}^{(2)}$ y $\mathbf{x}^{(1)}$ en lugar de $\mathbf{x}^{(1)}$ y $\mathbf{x}^{(0)}$.

En general, una vez que se ha determinado $\mathbf{x}^{(i)}$ se calcula $\mathbf{x}^{(i+1)}$ por medio de

$$A_i = A_{i-1} + \frac{\mathbf{y}_i - A_{i-1}\mathbf{s}_i}{||\mathbf{s}_i||_2^2}\mathbf{s}_i^t \tag{10.13}$$

y

$$\mathbf{x}^{(i+1)} = \mathbf{x}^{(i)} - A_i^{-1}\mathbf{F}\left(\mathbf{x}^{(i)}\right), \tag{10.14}$$

donde se introducen las notaciones $\mathbf{y}_i = \mathbf{F}\left(\mathbf{x}^{(i)}\right) - \mathbf{F}\left(\mathbf{x}^{(i-1)}\right)$ y $\mathbf{s}_i = \mathbf{x}^{(i)} - \mathbf{x}^{(i-1)}$ para simplificar las ecuaciones.

Si los métodos se realizan de acuerdo con lo descrito en las ecuaciones (10.13) y (10.14), el número de evaluaciones funcionales escalares se reduciría desde $n^2 + n$ hasta n (las requeridas para evaluar $\mathbf{F}\left(\mathbf{x}^{(i)}\right)$), pero se seguirían necesitando $O(n^3)$ cálculos para resolver el sistema lineal $n \times n$ (consulte el paso 4 en el algoritmo 10.1)

$$A_i\mathbf{s}_{i+1} = -\mathbf{F}\left(\mathbf{x}^{(i)}\right). \tag{10.15}$$

Esta forma de usar el método podría no estar justificada debido a la reducción de la convergencia superlineal a partir de la convergencia cuadrática del método de Newton.

Fórmula Sherman-Morrison

Sin embargo, es posible incorporar una mejora considerable al usar una fórmula de inversión de matriz de Sherman y Morrison (consulte, por ejemplo, [DM], p. 55). La prueba de esta fórmula se considera en los ejercicios 11 y 12.

Teorema 10.8 **(Fórmula Sherman-Morrison)**

Suponga que A es una matriz no singular y que $\mathbf{x}$ y $\mathbf{y}$ son los vectores con $\mathbf{y}^t A^{-1}\mathbf{x} \neq -1$. Entonces $A + \mathbf{x}\mathbf{y}^t$ es no singular y

$$\left(A + \mathbf{x}\mathbf{y}^t\right)^{-1} = A^{-1} - \frac{A^{-1}\mathbf{x}\mathbf{y}^t A^{-1}}{1 + \mathbf{y}^t A^{-1}\mathbf{x}}.$$ ■

La fórmula Sherman-Morrison permite calcular directamente A_i^{-1} a partir de A_{i-1}^{-1}, lo cual elimina la necesidad de una inversión de matriz con cada iteración.

Si hacemos $A = A_{i-1}, \mathbf{x} = (\mathbf{y}_i - A_{i-1}\mathbf{s}_i)/||\mathbf{s}_i||_2^2$, y $\mathbf{y} = \mathbf{s}_i$, en la ecuación (10.13) obtenemos

$$\begin{aligned}
A_i^{-1} &= \left(A_{i-1} + \frac{\mathbf{y}_i - A_{i-1}\mathbf{s}_i}{||\mathbf{s}_i||_2^2}\mathbf{s}_i^t\right)^{-1} \\
&= A_{i-1}^{-1} - \frac{A_{i-1}^{-1}\left(\frac{\mathbf{y}_i - A_{i-1}\mathbf{s}_i}{||\mathbf{s}_i||_2^2}\mathbf{s}_i^t\right)A_{i-1}^{-1}}{1 + \mathbf{s}_i^t A_{i-1}^{-1}\left(\frac{\mathbf{y}_i - A_{i-1}\mathbf{s}_i}{||\mathbf{s}_i||_2^2}\right)} \\
&= A_{i-1}^{-1} - \frac{\left(A_{i-1}^{-1}y_i - \mathbf{s}_i\right)\mathbf{s}_i^t A_{i-1}^{-1}}{||\mathbf{s}_i||_2^2 + \mathbf{s}_i^t A_{i-1}^{-1}\mathbf{y}_i - ||\mathbf{s}_i||_2^2},
\end{aligned}$$

por lo que

$$A_i^{-1} = A_{i-1}^{-1} + \frac{\left(\mathbf{s}_i - A_{i-1}^{-1}\mathbf{y}_i\right)\mathbf{s}_i^t A_{i-1}^{-1}}{\mathbf{s}_i^t A_{i-1}^{-1}\mathbf{y}_i}. \tag{10.16}$$

Este cálculo sólo implica multiplicaciones matriz-vector en cada paso y, por lo tanto, solamente requiere $O(n^2)$ cálculos aritméticos. Se evita el cálculo de A_i, al igual que la necesidad de resolver el sistema lineal (10.15).

El algoritmo 10.2 sigue directamente esta construcción, al incorporar la ecuación (10.16) en la técnica iterativa (10.14).

ALGORITMO 10.2

Método de Broyden

Para aproximar la solución del sistema no lineal $\mathbf{F}(\mathbf{x}) = \mathbf{0}$, dada una aproximación inicial $\mathbf{x}$:

ENTRADA número n de ecuaciones y variables; aproximación inicial $\mathbf{x} = (x_1, \ldots, x_n)^t$; tolerancia TOL; número máximo de iteraciones N.

SALIDA solución aproximada $\mathbf{x} = (x_1, \ldots, x_n)^t$ o un mensaje que indica que el número de iteraciones fue excedido.

Paso 1 Determine $A_0 = J(\mathbf{x})$ donde $J(\mathbf{x})_{i,j} = \frac{\partial f_i}{\partial x_j}(\mathbf{x})$ para $1 \leq i,\ j \leq n$;
$\mathbf{v} = \mathbf{F}(\mathbf{x})$. (*Nota:* $\mathbf{v} = \mathbf{F}(\mathbf{x}^{(0)})$.)

Paso 2 Determine $A = A_0^{-1}$. (*Use la eliminación gaussiana*.)

Paso 3 Determine $\mathbf{s} = -A\mathbf{v}$; (*Nota:* $\mathbf{s} = \mathbf{s}_1$.)
$\mathbf{x} = \mathbf{x} + \mathbf{s}$; (*Nota:* $\mathbf{x} = \mathbf{x}^{(1)}$.)
$k = 2$.

Paso 4 Mientras $(k \leq N)$ haga los pasos 5–13.

Paso 5 Determine $\mathbf{w} = \mathbf{v}$; (*Conserve* $\mathbf{v}$.)
$\mathbf{v} = \mathbf{F}(\mathbf{x})$; (*Nota:* $\mathbf{v} = \mathbf{F}(\mathbf{x}^{(k)})$.)
$\mathbf{y} = \mathbf{v} - \mathbf{w}$. (*Nota:* $\mathbf{y} = \mathbf{y}_k$.)

Paso 6 Determine $\mathbf{z} = -A\mathbf{y}$. (*Nota:* $\mathbf{z} = -A_{k-1}^{-1}\mathbf{y}_k$.)

Paso 7 Determine $p = -\mathbf{s}^t\mathbf{z}$. (*Nota:* $p = \mathbf{s}_k^t A_{k-1}^{-1}\mathbf{y}_k$.)

Paso 8 Determine $\mathbf{u}^t = \mathbf{s}^t A$.

Paso 9 Determine $A = A + \frac{1}{p}(\mathbf{s} + \mathbf{z})\mathbf{u}^t$. (*Nota:* $A = A_k^{-1}$.)

Paso 10 Determine $\mathbf{s} = -A\mathbf{v}$. (*Nota:* $\mathbf{s} = -A_k^{-1}\mathbf{F}(\mathbf{x}^{(k)})$.)

Paso 11 Determine $\mathbf{x} = \mathbf{x} + \mathbf{s}$. (*Nota:* $\mathbf{x} = \mathbf{x}^{(k+1)}$.)

Paso 12 Si $||\mathbf{s}|| < TOL$ entonces SALIDA ($\mathbf{x}$);
(*El procedimiento fue exitoso*.)
PARE.

Paso 13 Determine $k = k + 1$.

Paso 14 SALIDA ('Número máximo de iteraciones excedido');
(*El procedimiento no fue exitoso*.)
PARE. ■

Ejemplo 1 Use el método de Broyden con $\mathbf{x}^{(0)} = (0.1, 0.1, -0.1)^t$ para aproximar la solución del sistema no lineal

$$3x_1 - \cos(x_2x_3) - \frac{1}{2} = 0,$$

$$x_1^2 - 81(x_2 + 0.1)^2 + \operatorname{sen} x_3 + 1.06 = 0,$$

$$e^{-x_1x_2} + 20x_3 + \frac{10\pi - 3}{3} = 0.$$

Solución Este sistema se resolvió mediante el método de Newton en el ejemplo 1 de la sección 10.2. La matriz jacobiana para este sistema es

$$J(x_1, x_2, x_3) = \begin{bmatrix} 3 & x_3 \operatorname{sen} x_2x_3 & x_2 \operatorname{sen} x_2x_3 \\ 2x_1 & -162(x_2 + 0.1) & \cos x_3 \\ -x_2e^{-x_1x_2} & -x_1e^{-x_1x_2} & 20 \end{bmatrix}.$$

Si $\mathbf{x}^{(0)} = (0.1, 0.1, -0.1)^t$ y

$$\mathbf{F}(x_1, x_2, x_3) = (f_1(x_1, x_2, x_3),\ f_2(x_1, x_2, x_3),\ f_3(x_1, x_2, x_3))^t,$$

donde

$$f_1(x_1, x_2, x_3) = 3x_1 - \cos(x_2 x_3) - \frac{1}{2},$$

$$f_2(x_1, x_2, x_3) = x_1^2 - 81(x_2 + 0.1)^2 + \operatorname{sen} x_3 + 1.06,$$

y

$$f_3(x_1, x_2, x_3) = e^{-x_1 x_2} + 20x_3 + \frac{10\pi - 3}{3}.$$

Entonces

$$\mathbf{F}\left(\mathbf{x}^{(0)}\right) = \begin{bmatrix} -1.199950 \\ -2.269833 \\ 8.462025 \end{bmatrix}.$$

Puesto que

$$\begin{aligned} A_0 &= J\left(x_1^{(0)}, x_2^{(0)}, x_3^{(0)}\right) \\ &= \begin{bmatrix} 3 & 9.999833 \times 10^{-4} & -9.999833 \times 10^{-4} \\ 0.2 & -32.4 & 0.9950042 \\ -9.900498 \times 10^{-2} & -9.900498 \times 10^{-2} & 20 \end{bmatrix}, \end{aligned}$$

tenemos

$$\begin{aligned} A_0^{-1} &= J\left(x_1^{(0)}, x_2^{(0)}, x_3^{(0)}\right)^{-1} \\ &= \begin{bmatrix} 0.3333332 & 1.023852 \times 10^{-5} & 1.615701 \times 10^{-5} \\ 2.108607 \times 10^{-3} & -3.086883 \times 10^{-2} & 1.535836 \times 10^{-3} \\ 1.660520 \times 10^{-3} & -1.527577 \times 10^{-4} & 5.000768 \times 10^{-2} \end{bmatrix}. \end{aligned}$$

Por lo tanto

$$\mathbf{x}^{(1)} = \mathbf{x}^{(0)} - A_0^{-1}\mathbf{F}\left(\mathbf{x}^{(0)}\right) = \begin{bmatrix} 0.4998697 \\ 1.946685 \times 10^{-2} \\ -0.5215205 \end{bmatrix},$$

$$\mathbf{F}\left(\mathbf{x}^{(1)}\right) = \begin{bmatrix} -3.394465 \times 10^{-4} \\ -0.3443879 \\ 3.188238 \times 10^{-2} \end{bmatrix},$$

$$\mathbf{y}_1 = \mathbf{F}\left(\mathbf{x}^{(1)}\right) - \mathbf{F}\left(\mathbf{x}^{(0)}\right) = \begin{bmatrix} 1.199611 \\ 1.925445 \\ -8.430143 \end{bmatrix},$$

$$\mathbf{s}_1 = \begin{bmatrix} 0.3998697 \\ -8.053315 \times 10^{-2} \\ -0.4215204 \end{bmatrix},$$

$$\mathbf{s}_1^t A_0^{-1} \mathbf{y}_1 = 0.3424604,$$

$$\begin{aligned} A_1^{-1} &= A_0^{-1} + (1/0.3424604)\left[\left(\mathbf{s}_1 - A_0^{-1}\mathbf{y}_1\right)\mathbf{s}_1^t A_0^{-1}\right] \\ &= \begin{bmatrix} 0.3333781 & 1.11050 \times 10^{-5} & 8.967344 \times 10^{-6} \\ -2.021270 \times 10^{-3} & -3.094849 \times 10^{-2} & 2.196906 \times 10^{-3} \\ 1.022214 \times 10^{-3} & -1.650709 \times 10^{-4} & 5.010986 \times 10^{-2} \end{bmatrix}, \end{aligned}$$

y

$$\mathbf{x}^{(2)} = \mathbf{x}^{(1)} - A_1^{-1}\mathbf{F}\left(\mathbf{x}^{(1)}\right) = \begin{bmatrix} 0.4999863 \\ 8.737833 \times 10^{-3} \\ -0.5231746 \end{bmatrix}.$$

Las iteraciones adicionales se listan en la tabla 10.4. La quinta iteración del método de Broyden es ligeramente menos precisa que la cuarta iteración del método de Newton, dada en el ejemplo al final de la sección anterior. ■

Tabla 10.4

k	$x_1^{(k)}$	$x_2^{(k)}$	$x_3^{(k)}$	$\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\|_2$
3	0.5000066	8.672157×10^{-4}	−0.5236918	7.88×10^{-3}
4	0.5000003	6.083352×10^{-5}	−0.5235954	8.12×10^{-4}
5	0.5000000	-1.448889×10^{-6}	−0.5235989	6.24×10^{-5}
6	0.5000000	6.059030×10^{-9}	−0.5235988	1.50×10^{-6}

Los procedimientos también están disponibles de tal forma que se mantiene la convergencia, pero reducen significativamente el número de evaluaciones funcionales requeridas. Los métodos de este tipo fueron propuestos originalmente por Brown [Brow, K]. Una reseña y comparación de algunos métodos de este tipo que se usan de manera común puede encontrarse en [MC]. No obstante, en general, estos métodos son mucho más difíciles de implementar de manera eficiente que el método de Broyden.

La sección Conjunto de ejercicios 10.3 está disponible en línea. Encuentre la ruta de acceso en las páginas preliminares.

10.4 Técnicas de descenso más rápido

La ventaja de los métodos de Newton y cuasi-Newton para resolver sistemas de ecuaciones no lineales es su velocidad de convergencia una vez que se conoce una aproximación suficientemente exacta. Una debilidad de estos métodos es que se necesita una aproximación inicial precisa a la solución para garantizar la convergencia. El **método de descenso** más rápido que se considera en esta sección sólo converge linealmente para la solución, pero normalmente también convergerá para aproximaciones iniciales pobres. Por consiguiente, este método se usa para encontrar aproximaciones iniciales suficientemente exactas para las técnicas con base en Newton de la misma forma que se usa el método de bisección para una sola ecuación.

El nombre del método de descenso más rápido sigue de la aplicación tridimensional de señalamiento en la dirección descendente más rápida.

El método de descenso más rápido determina un mínimo local para una función multivariable de la forma $g : \mathbb{R}^n \to \mathbb{R}$. El método es valioso más allá de la aplicación como método de inicio para resolver sistemas no lineales. (En los ejercicios se consideran algunas otras aplicaciones.)

La conexión entre la minimización de una función de $\mathbb{R}^n$ a $\mathbb{R}$ y la solución de un sistema de ecuaciones no lineales se debe al hecho de que un sistema de la forma

$$\begin{aligned} f_1(x_1, x_2, \ldots, x_n) &= 0, \\ f_2(x_1, x_2, \ldots, x_n) &= 0, \\ \vdots \qquad & \quad \vdots \\ f_n(x_1, x_2, \ldots, x_n) &= 0, \end{aligned}$$

tiene una solución en $\mathbf{x} = (x_1, x_2, \ldots, x_n)^t$ precisamente cuando la función g definida por

$$g(x_1, x_2, \ldots, x_n) = \sum_{i=1}^{n} [f_i(x_1, x_2, \ldots, x_n)]^2$$

tiene el valor mínimo 0.

El método del descenso más rápido para encontrar un mínimo local para una función arbitraria g de $\mathbb{R}^n$ a $\mathbb{R}$ se puede describir de modo intuitivo de acuerdo con lo siguiente:

1. Evalúe g en una aproximación inicial $\mathbf{x}^{(0)} = \left(x_1^{(0)}, x_2^{(0)}, \ldots, x_n^{(0)}\right)^t$.
2. Determine una dirección desde $\mathbf{x}^{(0)}$ que resulte en un decremento del valor de g.
3. Mueva una cantidad adecuada en esta dirección y llame al nuevo valor $\mathbf{x}^{(1)}$.
4. Repita los pasos 1 a 3 con $\mathbf{x}^{(0)}$ reemplazado por $\mathbf{x}^{(1)}$.

El gradiente de una función

Antes de describir cómo seleccionar la dirección correcta y la distancia adecuada para moverse hacia esta dirección, necesitamos revisar algunos resultados a partir del cálculo. El teorema del valor extremo 1.9 establece que una función de una sola variable diferenciable puede tener un mínimo relativo sólo cuando su derivada es cero. Para ampliar este resultado para funciones multivariables, necesitamos la siguiente definición.

Definición 10.9 Para $g : \mathbb{R}^n \to \mathbb{R}$ el **gradiente** de g en $\mathbf{x} = (x_1, x_2, \ldots, x_n)^t$ se denota $\nabla g(\mathbf{x})$ y se define como

$$\nabla g(\mathbf{x}) = \left(\frac{\partial g}{\partial x_1}(\mathbf{x}), \frac{\partial g}{\partial x_2}(\mathbf{x}), \ldots, \frac{\partial g}{\partial x_n}(\mathbf{x})\right)^t. \quad \blacksquare$$

El gradiente proviene de la raíz de la palabra latina *gradi*, que significa "caminar". En este sentido, el gradiente de una superficie es la velocidad con la que "camina cuesta arriba"

El gradiente de una función multivariable es análogo a la derivada de una función de una sola variable en el sentido que una función multivariable diferenciable puede tener un mínimo relativo en $\mathbf{x}$ sólo cuando el gradiente en $\mathbf{x}$ es el vector cero. El gradiente tiene otra propiedad importante conectado con la minimización de funciones multivariable. Suponga que $\mathbf{v} = (v_1, v_2, \ldots, v_n)^t$ es un vector unitario en $\mathbb{R}^n$; es decir,

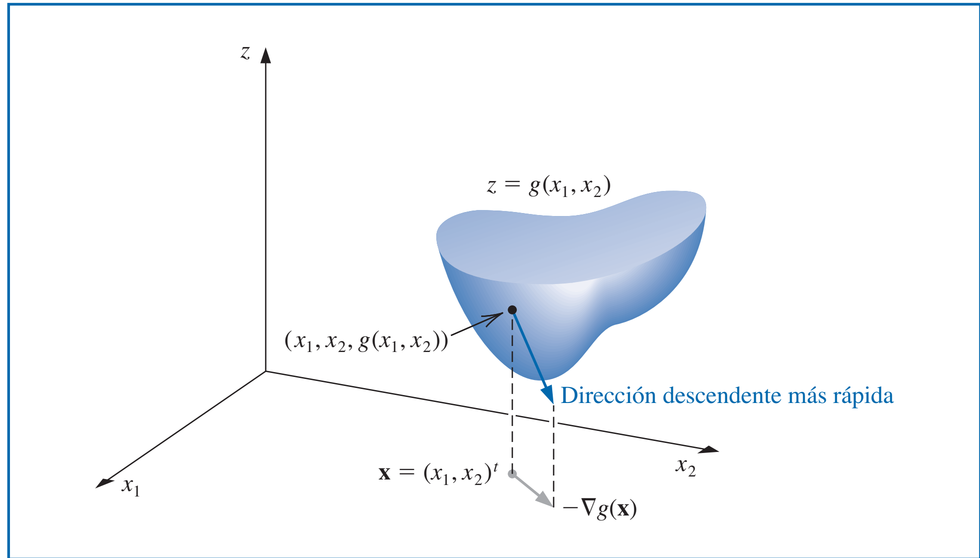
$$||\mathbf{v}||_2^2 = \sum_{i=1}^{n} v_i^2 = 1.$$

La **derivada direccional** de g en $\mathbf{x}$ en la dirección de $\mathbf{v}$ mide el cambio en el valor de la función g relativo al cambio en la variable en la dirección de $\mathbf{v}$. Se define mediante

$$D_{\mathbf{v}} g(\mathbf{x}) = \lim_{h\to 0} \frac{1}{h}[g(\mathbf{x} + h\mathbf{v}) - g(\mathbf{x})] = \mathbf{v}^t \cdot \nabla g(\mathbf{x}).$$

Cuando g es diferenciable, la dirección que produce el valor máximo para la derivada direccional se presenta cuando se selecciona $\mathbf{v}$ de forma paralela a $\nabla g(\mathbf{x})$, siempre y cuando $\nabla g(\mathbf{x}) \neq \mathbf{0}$. Por consiguiente, la dirección del mayor decremento en el valor de g en $\mathbf{x}$ es la dirección dada por $-\nabla g(\mathbf{x})$. La figura 10.2 es una ilustración cuando g es una función de dos variables.

Figura 10.2



El objetivo es reducir $g(\mathbf{x})$ hasta su valor mínimo de cero, por lo que una selección adecuada para $\mathbf{x}^{(1)}$ es apartarse de $\mathbf{x}^{(0)}$ en la dirección que proporciona el mayor descenso en el valor de $g(\mathbf{x})$. Por lo tanto, hacemos

$$\mathbf{x}^{(1)} = \mathbf{x}^{(0)} - \alpha \nabla g\left(\mathbf{x}^{(0)}\right), \quad \text{para alguna constante } \alpha > 0 \tag{10.17}$$

Ahora, el problema se reduce a seleccionar un valor adecuado de α por lo que $g(\mathbf{x}^{(1)})$ sería significativamente menor a $g(\mathbf{x}^{(0)})$.

Para determinar una selección adecuada para el valor α, consideramos la función de una sola variable

$$h(\alpha) = g\left(\mathbf{x}^{(0)} - \alpha \nabla g\left(\mathbf{x}^{(0)}\right)\right). \tag{10.18}$$

El valor de α que minimiza h es el valor necesario para la ecuación (10.17).

Encontrar un valor mínimo para h directamente requeriría diferenciar h y, después, resolver un problema de localización de raíz para determinar los puntos críticos de h. En general, este procedimiento es demasiado costoso. Por el contrario, elegimos tres números $\alpha_1 < \alpha_2 < \alpha_3$ que, esperamos, estén cerca de donde se presenta el valor mínimo de $h(\alpha)$. Entonces construimos el polinomio cuadrático $P(x)$ que interpola h en α_1, α_2 y α_3. El mínimo del polinomio cuadrático se encuentra fácilmente de forma similar a la que se usa en el método Müller en la sección 2.6.

Definimos $\hat{\alpha}$ en $[\alpha_1, \alpha_3]$ de tal forma que $P(\hat{\alpha})$ es un mínimo en $[\alpha_1, \alpha_3]$ y usamos $P(\hat{\alpha})$ para aproximar el valor mínimo de $h(\alpha)$. Entonces, se usa $\hat{\alpha}$ en la determinación de la iteración nueva para aproximar el valor mínimo de g:

$$\mathbf{x}^{(1)} = \mathbf{x}^{(0)} - \hat{\alpha} \nabla g\left(\mathbf{x}^{(0)}\right).$$

Puesto que $g(\mathbf{x}^{(0)})$ está disponible, para minimizar el cálculo, primero seleccionamos $\alpha_1 = 0$. A continuación se encuentra un número α_3 con $h(\alpha_3) < h(\alpha_1)$. (Puesto que α_1 no minimiza h, este número α_3 existe). Finalmente, se selecciona α_2 que será $\alpha_3/2$.

El valor mínimo de P en $[\alpha_1, \alpha_3]$ se presenta ya sea en el único punto crítico de P o en el extremo derecho α_3 porque, por suposición, $P(\alpha_3) = h(\alpha_3) < h(\alpha_1) = P(\alpha_1)$. Puesto que $P(x)$ es un polinomio cuadrático, el punto crítico se puede encontrar al resolver la ecuación lineal.

Ejemplo 1 Utilice el método de descenso más rápido con $\mathbf{x}^{(0)} = (0,0,0)^t$ para encontrar una aproximación inicial razonable para la solución del sistema no lineal

$$f_1(x_1,x_2,x_3) = 3x_1 - \cos(x_2x_3) - \frac{1}{2} = 0,$$
$$f_2(x_1,x_2,x_3) = x_1^2 - 81(x_2+0.1)^2 + \operatorname{sen} x_3 + 1.06 = 0,$$
$$f_3(x_1,x_2,x_3) = e^{-x_1x_2} + 20x_3 + \frac{10\pi - 3}{3} = 0.$$

Solución Si $g(x_1,x_2,x_3) = [f_1(x_1,x_2,x_3)]^2 + [f_2(x_1,x_2,x_3)]^2 + [f_3(x_1,x_2,x_3)]^2$. Entonces

$$\begin{aligned}\nabla g(x_1,x_2,x_3) \equiv \nabla g(\mathbf{x}) &= \left(2f_1(\mathbf{x})\frac{\partial f_1}{\partial x_1}(\mathbf{x}) + 2f_2(\mathbf{x})\frac{\partial f_2}{\partial x_1}(\mathbf{x}) + 2f_3(\mathbf{x})\frac{\partial f_3}{\partial x_1}(\mathbf{x}),\right.\\ &\quad 2f_1(\mathbf{x})\frac{\partial f_1}{\partial x_2}(\mathbf{x}) + 2f_2(\mathbf{x})\frac{\partial f_2}{\partial x_2}(\mathbf{x}) + 2f_3(\mathbf{x})\frac{\partial f_3}{\partial x_2}(\mathbf{x}),\\ &\quad \left.2f_1(\mathbf{x})\frac{\partial f_1}{\partial x_3}(\mathbf{x}) + 2f_2(\mathbf{x})\frac{\partial f_2}{\partial x_3}(\mathbf{x}) + 2f_3(\mathbf{x})\frac{\partial f_3}{\partial x_3}(\mathbf{x})\right)\\ &= 2\mathbf{J}(\mathbf{x})^t\mathbf{F}(\mathbf{x}).\end{aligned}$$

Para $\mathbf{x}^{(0)} = (0,0,0)^t$, tenemos

$$g\left(\mathbf{x}^{(0)}\right) = 111.975 \quad \text{y} \quad z_0 = ||\nabla g\left(\mathbf{x}^{(0)}\right)||_2 = 419.554.$$

Sea

$$\mathbf{z} = \frac{1}{z_0}\nabla g\left(\mathbf{x}^{(0)}\right) = (-0.0214514, -0.0193062, 0.999583)^t.$$

Con $\alpha_1 = 0$, tenemos $g_1 = g\left(\mathbf{x}^{(0)} - \alpha_1\mathbf{z}\right) = g\left(\mathbf{x}^{(0)}\right) = 111.975$. De manera arbitraria hacemos que $\alpha_3 = 1$ de tal forma que

$$g_3 = g\left(\mathbf{x}^{(0)} - \alpha_3\mathbf{z}\right) = 93.5649.$$

Puesto que $g_3 < g_1$, aceptamos α_3 y establecemos $\alpha_2 = \alpha_3/2 = 0.5$. Por lo tanto,

$$g_2 = g\left(\mathbf{x}^{(0)} - \alpha_2\mathbf{z}\right) = 2.53557.$$

Ahora, encontramos el polinomio cuadrático que interpola los datos (0, 111.975), $(1, 93.5649)$ y $(0.5, 2.53557)$. Para este propósito, es más conveniente usar un polinomio de interpolación de diferencias divididas hacia adelante, que tiene la forma

$$P(\alpha) = g_1 + h_1\alpha + h_3\alpha(\alpha - \alpha_2).$$

Esto interpola

$$g\left(\mathbf{x}^{(0)} - \alpha\nabla g\left(\mathbf{x}^{(0)}\right)\right) = g\left(\mathbf{x}^{(0)} - \alpha\mathbf{z}\right)$$

$\alpha_1 = 0$, $\alpha_2 = 0.5$, y $\alpha_3 = 1$, como sigue:

$$\alpha_1 = 0, \qquad g_1 = 111.975,$$

$$\alpha_2 = 0.5, \quad g_2 = 2.53557, \quad h_1 = \frac{g_2 - g_1}{\alpha_2 - \alpha_1} = -218.878,$$

$$\alpha_3 = 1, \qquad g_3 = 93.5649, \quad h_2 = \frac{g_3 - g_2}{\alpha_3 - \alpha_2} = 182.059, \qquad h_3 = \frac{h_2 - h_1}{\alpha_3 - \alpha_1} = 400.937.$$

Por lo tanto,

$$P(\alpha) = 111.975 - 218.878\alpha + 400.937\alpha(\alpha - 0.5).$$

Tenemos $P'(\alpha) = 0$ cuando $\alpha = \alpha_0 = 0.522959$. Puesto que $g_0 = g\left(\mathbf{x}^{(0)} - \alpha_0 \mathbf{z}\right) = 2.32762$ es más pequeño que g_1 y g_3, establecemos

$$\mathbf{x}^{(1)} = \mathbf{x}^{(0)} - \alpha_0 \mathbf{z} = \mathbf{x}^{(0)} - 0.522959\mathbf{z} = (0.0112182, 0.0100964, -0.522741)^t$$

y

$$g\left(\mathbf{x}^{(1)}\right) = 2.32762.$$

La tabla 10.5 contiene el resto de los resultados. La verdadera solución para el sistema no lineal es $(0.5, 0, -0.5235988)^t$, por lo que $\mathbf{x}^{(2)}$ probablemente sería adecuado como una aproximación inicial para el método de Newton o de Broyden. Una de las técnicas que convergen más rápidamente serían adecuadas en esta etapa ya que se requieren 70 iteraciones del método de descenso más rápido para encontrar $\|\mathbf{x}^{(k)} - \mathbf{x}\|_\infty < 0.01$. ■

Tabla 10.5

k	$x_1^{(k)}$	$x_2^{(k)}$	$x_3^{(k)}$	$g(x_1^{(k)}, x_2^{(k)}, x_3^{(k)})$
2	0.137860	−0.205453	−0.522059	1.27406
3	0.266959	0.00551102	−0.558494	1.06813
4	0.272734	−0.00811751	−0.522006	0.468309
5	0.308689	−0.0204026	−0.533112	0.381087
6	0.314308	−0.0147046	−0.520923	0.318837
7	0.324267	−0.00852549	−0.528431	0.287024

El algoritmo 10.3 implica el método de descenso más rápido para aproximar el valor mínimo de $g(\mathbf{x})$. Para comenzar una iteración, el valor 0 se asigna a α_1, y el valor 1 se asigna a α_3. Si $h(\alpha_3) \geq h(\alpha_1)$, entonces se realizan las divisiones sucesivas de α_3 entre 2 y el valor de α_3 se reasigna hasta que $h(\alpha_3) < h(\alpha_1)$ y $\alpha_3 = 2^{-k}$ para algún valor de k.

Para usar el método para aproximar la solución para el sistema

$$\begin{aligned} f_1(x_1, x_2, \ldots, x_n) &= 0, \\ f_2(x_1, x_2, \ldots, x_n) &= 0, \\ \vdots \qquad &\quad \vdots \\ f_n(x_1, x_2, \ldots, x_n) &= 0, \end{aligned}$$

simplemente reemplazamos la función g con $\sum_{i=1}^{n} f_i^2$.

ALGORITMO **10.3**

Descenso más rápido

Para aproximar una solución $\mathbf{p}$ del problema de minimización

$$g(\mathbf{p}) = \min_{\mathbf{x} \in \mathbb{R}^n} g(\mathbf{x})$$

dada una aproximación inicial $\mathbf{x}$:

ENTRADA número n de variables; aproximación inicial $\mathbf{x} = (x_1, \ldots, x_n)^t$; tolerancia TOL; número máximo de iteraciones N.

SALIDA solución aproximada $\mathbf{x} = (x_1, \ldots, x_n)^t$ o un mensaje de falla.

Paso 1 Determine $k = 1$.

Paso 2 Mientras ($k \leq N$) haga los pasos 3–15.

Paso 3 Determine $g_1 = g(x_1, \dots, x_n)$; $\left(\textit{Nota: } g_1 = g\left(\mathbf{x}^{(k)}\right).\right)$
$\mathbf{z} = \nabla g(x_1, \dots, x_n)$; $\left(\textit{Nota: } \mathbf{z} = \nabla g\left(\mathbf{x}^{(k)}\right).\right)$
$z_0 = ||\mathbf{z}||_2$.

Paso 4 Si $z_0 = 0$ entonces SALIDA ('gradiente cero');
SALIDA ($x_1, \dots, x_n, g_1$);
(*Procedimiento completo, puede tener un mínimo*.)
PARE.

Paso 5 Determine $\mathbf{z} = \mathbf{z}/z_0$; (*Convierta* $\mathbf{z}$ *en vector unidad*.)
$\alpha_1 = 0$;
$\alpha_3 = 1$;
$g_3 = g(\mathbf{x} - \alpha_3\mathbf{z})$.

Paso 6 Mientras ($g_3 \geq g_1$) haga los pasos 7 y 8.

Paso 7 Determine $\alpha_3 = \alpha_3/2$;
$g_3 = g(\mathbf{x} - \alpha_3\mathbf{z})$.

Paso 8 Si $\alpha_3 < TOL/2$ entonces
SALIDA ('Sin probable mejora);
SALIDA ($x_1, \dots, x_n, g_1$);
(*Procedimiento completado, puede tener un mínimo*.)
PARE.

Paso 9 Determine $\alpha_2 = \alpha_3/2$;
$g_2 = g(\mathbf{x} - \alpha_2\mathbf{z})$.

Paso 10 Determine $h_1 = (g_2 - g_1)/\alpha_2$;
$h_2 = (g_3 - g_2)/(\alpha_3 - \alpha_2)$;
$h_3 = (h_2 - h_1)/\alpha_3$.
(*Nota: La fórmula de diferencias divididas hacia adelante de Newton se usa para encontrar* $P(\alpha) = g_1 + h_1\alpha + h_3\alpha(\alpha - \alpha_2)$ *la cuadrática que interpola* $h(\alpha)$ *en* $\alpha = 0, \alpha = \alpha_2, \alpha = \alpha_3$.)

Paso 11 Determine $\alpha_0 = 0.5(\alpha_2 - h_1/h_3)$; (*El punto crítico de P se presenta en* α_0.)
$g_0 = g(\mathbf{x} - \alpha_0\mathbf{z})$.

Paso 12 Encuentre α de $\{\alpha_0, \alpha_3\}$ de modo que $g = g(\mathbf{x} - \alpha\mathbf{z}) = \text{mín}\{g_0, g_3\}$.

Paso 13 Determine $\mathbf{x} = \mathbf{x} - \alpha\mathbf{z}$.

Paso 14 Si $|g - g_1| < TOL$ entonces
SALIDA ($x_1, \dots, x_n, g$);
(*El procedimiento fue exitoso*.)
PARE.

Paso 15 Determine $k = k + 1$.

Paso 16 SALIDA ('iteraciones máximas excedidas');
(*El procedimiento no fue exitoso*.)
PARE. ■

Existen muchas variaciones del método de descenso más rápido, algunos de los cuales implican métodos más complejos para determinar el valor de α que producirá un mínimo para la función de variable única h definida en la ecuación (10.18). Otras técnicas usan un polinomio de Taylor multidimensional para reemplazar la función multivariable g original y minimizar el polinomio en lugar de g. A pesar de que existen ventajas para algunos de estos métodos sobre el procedimiento analizado aquí, todos los métodos de descenso más rápido son, en general, linealmente convergentes y convergen independientemente de la aproximación inicial. En algunos casos, sin embargo, los métodos pueden converger en algo más que el mínimo absoluto de la función g.

Un análisis más completo de los métodos de descenso más rápido se puede encontrar en [OR] o [RR].

La sección Conjunto de ejercicios 10.4 está disponible en línea. Encuentre la ruta de acceso en las páginas preliminares.

10.5 Homotopía y métodos de continuación

Los métodos de *homotopía*, o *continuación*, para los sistemas no lineales insertaron el problema que debe resolverse dentro de un conjunto de problemas. En específico, para resolver un problema de la forma

$$\mathbf{F}(\mathbf{x}) = \mathbf{0},$$

que posee la solución desconocida $\mathbf{x}^*$, consideramos una familia de problemas descritos usando un parámetro λ que supone valores en [0, 1]. Un problema con una solución conocida $\mathbf{x}(0)$ corresponde a la situación cuando $\lambda = 0$ y el problema con la solución desconocida $\mathbf{x}(1) \equiv \mathbf{x}^*$ corresponde a $\lambda = 1$.

Una homotopía es una deformación continua, una función que toma un intervalo real continuamente en un conjunto de funciones.

Por ejemplo suponga que $\mathbf{x}(0)$ es una aproximación inicial para la solución de $\mathbf{F}(\mathbf{x}^*) = \mathbf{0}$. Defina

$$\mathbf{G} : [0, 1] \times \mathbb{R}^n \to \mathbb{R}^n$$

mediante

$$\mathbf{G}(\lambda, \mathbf{x}) = \lambda\mathbf{F}(\mathbf{x}) + (1 - \lambda)\,[\mathbf{F}(\mathbf{x}) - \mathbf{F}(\mathbf{x}(0))] = \mathbf{F}(\mathbf{x}) + (\lambda - 1)\mathbf{F}(\mathbf{x}(0)). \quad (10.19)$$

Determinaremos, para varios valores de λ, una solución para

$$\mathbf{G}(\lambda, \mathbf{x}) = \mathbf{0}.$$

Cuando $\lambda = 0$, esta ecuación asume la forma

$$\mathbf{0} = \mathbf{G}(0, \mathbf{x}) = \mathbf{F}(\mathbf{x}) - \mathbf{F}(\mathbf{x}(0)),$$

y $\mathbf{x}(0)$ es una solución. Cuando $\lambda = 1$, la ecuación asume la forma

$$\mathbf{0} = \mathbf{G}(1, \mathbf{x}) = \mathbf{F}(\mathbf{x}),$$

y $\mathbf{x}(1) = \mathbf{x}^*$ es una solución.

La función $\mathbf{G}$, con el parámetro λ, nos proporciona una familia de funciones que puede conducir al valor conocido $\mathbf{x}(0)$ para la solución $\mathbf{x}(1) = \mathbf{x}^*$. La función $\mathbf{G}$ recibe el nombre de **homotopía** entre la función $\mathbf{G}(0, \mathbf{x}) = \mathbf{F}(\mathbf{x}) - \mathbf{F}(\mathbf{x}(0))$ y la función $\mathbf{G}(1, \mathbf{x}) = \mathbf{F}(\mathbf{x})$.

Método de continuación

El problema de **continuación** es:

- Determinar una forma de proceder con la solución conocida $\mathbf{x}(0)$ de $\mathbf{G}(0, \mathbf{x}) = \mathbf{0}$ para la solución desconocida $\mathbf{x}(1) = \mathbf{x}^*$ de $\mathbf{G}(1, \mathbf{x}) = \mathbf{0}$, es decir, la solución para $\mathbf{F}(\mathbf{x}) = \mathbf{0}$.

Primero suponemos que $\mathbf{x}(\lambda)$ es una solución única para la ecuación

$$\mathbf{G}(\lambda, \mathbf{x}) = \mathbf{0}, \tag{10.20}$$

para cada $\lambda \in [0, 1]$. El conjunto $\{\mathbf{x}(\lambda) \mid 0 \le \lambda \le 1\}$ se puede observar como una curva en $\mathbb{R}^n$ desde $\mathbf{x}(0)$ hasta $\mathbf{x}(1) = \mathbf{x}^*$ parametrizada por λ. Un método de continuación encuentra una sucesión de pasos a lo largo de esta curva correspondiente a $\{\mathbf{x}(\lambda_k)\}_{k=0}^{m}$, donde $\lambda_0 = 0 < \lambda_1 < \cdots < \lambda_m = 1$.

Si las funciones $\lambda \to \mathbf{x}(\lambda)$ y $\mathbf{G}$ son diferenciables, entonces derivar la ecuación (10.20) respecto a λ nos da

$$\mathbf{0} = \frac{\partial \mathbf{G}(\lambda, \mathbf{x}(\lambda))}{\partial \lambda} + \frac{\partial \mathbf{G}(\lambda, \mathbf{x}(\lambda))}{\partial \mathbf{x}} \mathbf{x}'(\lambda),$$

y resolver para $\mathbf{x}'(\lambda)$ nos da

$$\mathbf{x}'(\lambda) = -\left[\frac{\partial \mathbf{G}(\lambda, \mathbf{x}(\lambda))}{\partial \mathbf{x}}\right]^{-1} \frac{\partial \mathbf{G}(\lambda, \mathbf{x}(\lambda))}{\partial \lambda}.$$

Éste es un sistema de ecuaciones diferenciales con la condición inicial $\mathbf{x}(0)$.

Puesto que

$$\mathbf{G}(\lambda, \mathbf{x}(\lambda)) = \mathbf{F}(\mathbf{x}(\lambda)) + (\lambda - 1)\mathbf{F}(\mathbf{x}(0)),$$

podemos determinar tanto

$$\frac{\partial \mathbf{G}}{\partial \mathbf{x}}(\lambda, \mathbf{x}(\lambda)) = \begin{bmatrix} \frac{\partial f_1}{\partial x_1}(\mathbf{x}(\lambda)) & \frac{\partial f_1}{\partial x_2}(\mathbf{x}(\lambda)) & \dots & \frac{\partial f_1}{\partial x_n}(\mathbf{x}(\lambda)) \\ \frac{\partial f_2}{\partial x_1}(\mathbf{x}(\lambda)) & \frac{\partial f_2}{\partial x_2}(\mathbf{x}(\lambda)) & \dots & \frac{\partial f_2}{\partial x_n}(\mathbf{x}(\lambda)) \\ \vdots & \vdots & & \vdots \\ \frac{\partial f_n}{\partial x_1}(\mathbf{x}(\lambda)) & \frac{\partial f_n}{\partial x_2}(\mathbf{x}(\lambda)) & \dots & \frac{\partial f_n}{\partial x_n}(\mathbf{x}(\lambda)) \end{bmatrix} = J(\mathbf{x}(\lambda)),$$

la matriz jacobiana como

$$\frac{\partial \mathbf{G}(\lambda, \mathbf{x}(\lambda))}{\partial \lambda} = \mathbf{F}(\mathbf{x}(0)).$$

Por lo tanto, el sistema de ecuaciones diferenciales se convierte en

$$\mathbf{x}'(\lambda) = -[J(\mathbf{x}(\lambda))]^{-1}\mathbf{F}(\mathbf{x}(0)), \quad \text{para} \quad 0 \le \lambda \le 1, \tag{10.21}$$

con la condición inicial $\mathbf{x}(0)$. El siguiente teorema (consulte [OR], pp. 230–231) da las condiciones en las que el método de continuación es factible.

Teorema 10.10 Sea $\mathbf{F}(\mathbf{x})$ continuamente diferenciable en $\mathbf{x} \in \mathbb{R}^n$. Suponga que la matriz jacobiana $J(\mathbf{x})$ es no singular para todas las $\mathbf{x} \in \mathbb{R}^n$ y que la constante M existe con $\|J(\mathbf{x})^{-1}\| \le M$, para todas las $\mathbf{x} \in \mathbb{R}^n$. Entonces, para cualquier $\mathbf{x}(0)$ en $\mathbb{R}^n$, existe una única función $\mathbf{x}(\lambda)$, tal que

$$\mathbf{G}(\lambda, \mathbf{x}(\lambda)) = \mathbf{0},$$

para toda λ en [0, 1]. Además, $\mathbf{x}(\lambda)$ es continuamente diferenciable y

$$\mathbf{x}'(\lambda) = -J(\mathbf{x}(\lambda))^{-1}\mathbf{F}(\mathbf{x}(0)), \quad \text{para cada } \lambda \in [0, 1]. \quad ■$$

Lo siguiente muestra la forma del sistema de ecuaciones diferenciales relacionadas con un sistema de ecuaciones no lineales.

Ilustración Considere el sistema no lineal

$$f_1(x_1, x_2, x_3) = 3x_1 - \cos(x_2 x_3) - 0.5 = 0,$$
$$f_2(x_1, x_2, x_3) = x_1^2 - 81(x_2 + 0.1)^2 + \operatorname{sen} x_3 + 1.06 = 0,$$
$$f_3(x_1, x_2, x_3) = e^{-x_1 x_2} + 20x_3 + \frac{10\pi - 3}{3} = 0.$$

La matriz jacobiana es

$$J(\mathbf{x}) = \begin{bmatrix} 3 & x_3 \operatorname{sen} x_2 x_3 & x_2 \operatorname{sen} x_2 x_3 \\ 2x_1 & -162(x_2 + 0.1) & \cos x_3 \\ -x_2 e^{-x_1 x_2} & -x_1 e^{-x_1 x_2} & 20 \end{bmatrix}.$$

Sea $\mathbf{x}(0) = (0, 0, 0)^t$, así que

$$\mathbf{F}(\mathbf{x}(0)) = \begin{bmatrix} -1.5 \\ 0.25 \\ 10\pi/3 \end{bmatrix}.$$

El sistema de ecuaciones diferenciales es

$$\begin{bmatrix} x_1'(\lambda) \\ x_2'(\lambda) \\ x_3'(\lambda) \end{bmatrix} = -\begin{bmatrix} 3 & x_3 \operatorname{sen} x_2 x_3 & x_2 \operatorname{sen} x_2 x_3 \\ 2x_1 & -162(x_2 + 0.1) & \cos x_3 \\ -x_2 e^{-x_1 x_2} & -x_1 e^{-x_1 x_2} & 20 \end{bmatrix}^{-1} \begin{bmatrix} -1.5 \\ 0.25 \\ 10\pi/3 \end{bmatrix}. \quad ■$$

En general, el sistema de ecuaciones diferenciales que necesitamos resolver para nuestro problema de continuación tiene la forma

$$\frac{dx_1}{d\lambda} = \phi_1(\lambda, x_1, x_2, \ldots, x_n),$$
$$\frac{dx_2}{d\lambda} = \phi_2(\lambda, x_1, x_2, \ldots, x_n),$$
$$\vdots$$
$$\frac{dx_n}{d\lambda} = \phi_n(\lambda, x_1, x_2, \ldots, x_n),$$

donde

$$\begin{bmatrix} \phi_1(\lambda, x_1, \ldots, x_n) \\ \phi_2(\lambda, x_1, \ldots, x_n) \\ \vdots \\ \phi_n(\lambda, x_1, \ldots, x_n) \end{bmatrix} = -J(x_1, \ldots, x_n)^{-1} \begin{bmatrix} f_1(\mathbf{x}(0)) \\ f_2(\mathbf{x}(0)) \\ \vdots \\ f_n(\mathbf{x}(0)) \end{bmatrix}. \tag{10.22}$$

Para usar método Runge-Kutta de orden 4 en la resolución de este sistema, primero seleccionamos un entero $N > 0$ y establecemos que $h = (1 - 0)/N$. La partición del intervalo [0, 1] en N subintervalos con los puntos de malla

$$\lambda_j = jh, \quad \text{para cada } j = 0, 1, \ldots, N.$$

Usamos la notación w_{ij} para cada $j = 0, 1, \ldots, N$ y $i = 1, \ldots, n$, para denotar una aproximación para $x_i(\lambda_j)$. Para las condiciones iniciales, establecemos

$$w_{1,0} = x_1(0), \quad w_{2,0} = x_2(0), \quad \ldots, \quad w_{n,0} = x_n(0).$$

Suponga que se ha calculado $w_{1,j}, w_{2,j}, \ldots, w_{n,j}$. Obtenemos $w_{1,j+1}, w_{2,j+1}, \ldots, w_{n,j+1}$ usando las ecuaciones

$$k_{1,i} = h\phi_i(\lambda_j, w_{1,j}, w_{2,j}, \ldots, w_{n,j}), \quad \text{para cada } i = 1, 2, \ldots, n;$$

$$k_{2,i} = h\phi_i\left(\lambda_j + \frac{h}{2}, w_{1,j} + \frac{1}{2}k_{1,1}, \ldots, w_{n,j} + \frac{1}{2}k_{1,n}\right), \quad \text{para cada } i = 1, 2, \ldots, n;$$

$$k_{3,i} = h\phi_i\left(\lambda_j + \frac{h}{2}, w_{1,j} + \frac{1}{2}k_{2,1}, \ldots, w_{n,j} + \frac{1}{2}k_{2,n}\right), \quad \text{para cada } i = 1, 2, \ldots, n;$$

$$k_{4,i} = h\phi_i(\lambda_j + h, w_{1,j} + k_{3,1}, w_{2,j} + k_{3,2}, \ldots, w_{n,j} + k_{3,n}), \quad \text{para cada } i = 1, 2, \ldots, n;$$

y, finalmente,

$$w_{i,j+1} = w_{i,j} + \frac{1}{6}\left(k_{1,i} + 2k_{2,i} + 2k_{3,i} + k_{4,i}\right), \quad \text{para cada } i = 1, 2, \ldots, n.$$

La notación de vector

$$\mathbf{k}_1 = \begin{bmatrix} k_{1,1} \\ k_{1,2} \\ \vdots \\ k_{1,n} \end{bmatrix}, \ \mathbf{k}_2 = \begin{bmatrix} k_{2,1} \\ k_{2,2} \\ \vdots \\ k_{2,n} \end{bmatrix}, \ \mathbf{k}_3 = \begin{bmatrix} k_{3,1} \\ k_{3,2} \\ \vdots \\ k_{3,n} \end{bmatrix}, \ \mathbf{k}_4 = \begin{bmatrix} k_{4,1} \\ k_{4,2} \\ \vdots \\ k_{4,n} \end{bmatrix}, \quad \text{y} \quad \mathbf{w}_j = \begin{bmatrix} w_{1,j} \\ w_{2,j} \\ \vdots \\ w_{n,j} \end{bmatrix}$$

simplifica la presentación. Entonces la ecuación (10.22) nos da $\mathbf{x}(0) = \mathbf{x}(\lambda_0) = \mathbf{w}_0$, y por cada $j = 0, 1, \ldots, N$,

$$\begin{aligned}
\mathbf{k}_1 &= h \begin{bmatrix} \phi_1(\lambda_j, w_{1,j}, \ldots, w_{n,j}) \\ \phi_2(\lambda_j, w_{1,j}, \ldots, w_{n,j}) \\ \vdots \\ \phi_n(\lambda_j, w_{1,j}, \ldots, w_{n,j}) \end{bmatrix} = h\left[-J(w_{1,j}, \ldots, w_{n,j})\right]^{-1}\mathbf{F}(\mathbf{x}(0)) \\
&= h\left[-J(\mathbf{w}_j)\right]^{-1}\mathbf{F}(\mathbf{x}(0)), \\
\mathbf{k}_2 &= h\left[-J\left(\mathbf{w}_j + \frac{1}{2}\mathbf{k}_1\right)\right]^{-1}\mathbf{F}(\mathbf{x}(0)), \\
\mathbf{k}_3 &= h\left[-J\left(\mathbf{w}_j + \frac{1}{2}\mathbf{k}_2\right)\right]^{-1}\mathbf{F}(\mathbf{x}(0)), \\
\mathbf{k}_4 &= h\left[-J\left(\mathbf{w}_j + \mathbf{k}_3\right)\right]^{-1}\mathbf{F}(\mathbf{x}(0)),
\end{aligned}$$

y

$$\mathbf{x}(\lambda_{j+1}) = \mathbf{x}(\lambda_j) + \frac{1}{6}\left(\mathbf{k}_1 + 2\mathbf{k}_2 + 2\mathbf{k}_3 + \mathbf{k}_4\right) = \mathbf{w}_j + \frac{1}{6}\left(\mathbf{k}_1 + 2\mathbf{k}_2 + 2\mathbf{k}_3 + \mathbf{k}_4\right).$$

Finalmente, $\mathbf{x}(\lambda_n) = \mathbf{x}(1)$ es nuestra aproximación para $\mathbf{x}^*$.

Ejemplo 1 Use el método de continuación con $\mathbf{x}(0) = (0, 0, 0)^t$ para aproximar la solución de

$$f_1(x_1, x_2, x_3) = 3x_1 - \cos(x_2 x_3) - 0.5 = 0,$$

$$f_2(x_1, x_2, x_3) = x_1^2 - 81(x_2 + 0.1)^2 + \operatorname{sen} x_3 + 1.06 = 0,$$

$$f_3(x_1, x_2, x_3) = e^{-x_1, x_2} + 20x_3 + \frac{10\pi - 3}{3} = 0.$$

Solución La matriz jacobiana es

$$J(\mathbf{x}) = \begin{bmatrix} 3 & x_3 \operatorname{sen} x_2 x_3 & x_2 \operatorname{sen} x_2 x_3 \\ 2x_1 & -162(x_2 + 0.1) & \cos x_3 \\ -x_2 e^{-x_1 x_2} & -x_1 e^{-x_1 x_2} & 20 \end{bmatrix}$$

y

$$F(\mathbf{x}(0)) = (-1.5, 0.25, 10\pi/3)^t.$$

Con $N = 4$ y $h = 0.25$, tenemos

$$\mathbf{k}_1 = h[-J(\mathbf{x}^{(0)})]^{-1} F(\mathbf{x}(0)) = 0.25 \begin{bmatrix} 3 & 0 & 0 \\ 0 & -16.2 & 1 \\ 0 & 0 & 20 \end{bmatrix}^{-1} \begin{bmatrix} -1.5 \\ 0.25 \\ 10\pi/3 \end{bmatrix}$$

$$= (0.125, -0.004222203325, -0.1308996939)^t,$$

$$\mathbf{k}_2 = h[-J(0.0625, -0.002111101663, -0.06544984695)]^{-1}(-1.5, 0.25, 10\pi/3)^t$$

$$= 0.25 \begin{bmatrix} 3 & -0.9043289149 \times 10^{-5} & -0.2916936196 \times 10^{-6} \\ 0.125 & -15.85800153 & 0.9978589232 \\ 0.002111380229 & -0.06250824706 & 20 \end{bmatrix}^{-1} \begin{bmatrix} -1.5 \\ 0.25 \\ 10\pi/3 \end{bmatrix}$$

$$= (0.1249999773, -0.003311761993, -0.1309232406)^t,$$

$$\mathbf{k}_3 = h[-J(0.06249998865, -0.001655880997, -0.0654616203)]^{-1}(-1.5, 0.25, 10\pi/3)^t$$

$$= (0.1249999844, -0.003296244825, -0.130920346)^t,$$

$$\mathbf{k}_4 = h[-J(0.1249999844, -0.003296244825, -0.130920346)]^{-1}(-1.5, 0.25, 10\pi/3)^t$$

$$= (0.1249998945, -0.00230206762, -0.1309346977)^t,$$

y

$$\mathbf{x}(\lambda_1) = \mathbf{w}_1 = \mathbf{w}_0 + \frac{1}{6}(\mathbf{k}_1 + 2\mathbf{k}_2 + 2\mathbf{k}_3 + \mathbf{k}_4)$$

$$= (0.1249999697, -0.00329004743, -0.1309202608)^t.$$

Al continuar, tenemos

$$\mathbf{x}(\lambda_2) = \mathbf{w}_2 = (0.2499997679, -0.004507400128, -0.2618557619)^t,$$

$$\mathbf{x}(\lambda_3) = \mathbf{w}_3 = (0.3749996956, -0.003430352103, -0.3927634423)^t,$$

y

$$\mathbf{x}(\lambda_4) = \mathbf{x}(1) = \mathbf{w}_4 = (0.4999999954, 0.126782 \times 10^{-7}, -0.5235987758)^t.$$

Estos resultados son muy precisos porque la solución real es $(0.5, 0, -0.52359877)^t$. ■

Observe que en el método de Runge-Kutta, los pasos similares a

$$\mathbf{k}_i = h[-J(\mathbf{x}(\lambda_i) + \alpha_{i-1}\mathbf{k}_{i-1})]^{-1}\mathbf{F}(\mathbf{x}(0))$$

se pueden escribir como solución para $\mathbf{k}_i$ en el sistema lineal

$$J(\mathbf{x}(\lambda_i) + \alpha_{i-1}\mathbf{k}_{i-1})\,\mathbf{k}_i = -h\mathbf{F}(\mathbf{x}(0)).$$

Por lo que en el método Runge-Kutta de orden 4, el cálculo de cada $\boldsymbol{w}_j$ requiere resolver cuatro sistemas lineales, cada uno al calcular $\mathbf{k}_1$, $\mathbf{k}_2$, $\mathbf{k}_3$ y $\mathbf{k}_4$. Por lo tanto, para usar N pasos se requiere resolver $4N$ sistemas lineales. Por comparación, el método de Newton requiere resolver un sistema lineal por iteración. Por tanto, el trabajo relacionado con el método de Runge-Kutta es aproximadamente equivalente a $4N$ iteraciones del método de Newton.

Una alternativa es usar el método de Runge-Kutta de orden 2, de tal forma que el método de Euler modificado o incluso el método de Euler, para disminuir el número de sistemas lineales que se tienen que resolver. Otra posibilidad es usar valores más pequeños de N. Lo siguiente ilustra estas ideas.

Ilustración La tabla 10.6 resume una comparación del método de Euler, el método de punto medio y el método de Runge-Kutta de orden 4 aplicado al problema en el ejemplo, con una aproximación inicial $\mathbf{x}(0) = (0, 0, 0)^t$. La columna a la derecha en la tabla enumera el número de sistemas lineales que se requieren para la solución. ■

Tabla 10.6

Método	N	$\mathbf{x}(1)$	Sistemas
Euler	1	$(0.5, -0.0168888133, -0.5235987755)^t$	1
Euler	4	$(0.499999379, -0.004309160698, -0.523679652)^t$	4
Punto medio	1	$(0.4999966628, -0.00040240435, -0.523815371)^t$	2
Punto medio	4	$(0.500000066, -0.00001760089, -0.5236127761)^t$	8
Runge-Kutta	1	$(0.4999989843, -0.1676151 \times 10^{-5}, -0.5235989561)^t$	4
Runge-Kutta	4	$(0.4999999954, 0.126782 \times 10^{-7}, -0.5235987758)^t$	16

El método de continuación se puede usar como un método independiente y no requiere una selección particularmente buena de $\mathbf{x}(0)$. Sin embargo, el método también puede usarse para proporcionar una aproximación inicial para los métodos de Newton o Broyden. Por ejemplo, el resultado obtenido en el ejemplo 2 usando el método de Euler y $N = 2$ podría ser fácilmente suficiente para iniciar los métodos de Newton o Broyden más eficientes y podrían ser mejores para este objetivo que los métodos de continuación, lo cual requiere más cálculos. El algoritmo 10.4 es una implementación del método de continuación.

ALGORITMO 10.4

Algoritmo de continuación

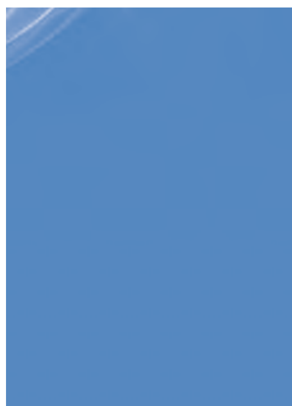
Para aproximar la solución del sistema no lineal $\mathbf{F}(\mathbf{x}) = \mathbf{0}$ dada una aproximación inicial $\mathbf{x}$:

ENTRADA número n de ecuaciones y de incógnitas; entero $N > 0$; aproximación inicial $\mathbf{x} = (x_1, x_2, \ldots, x_n)^t$.

SALIDA solución aproximada $\mathbf{x} = (x_1, x_2, \ldots, x_n)^t$.

Paso 1 Determine $h = 1/N$;
$\mathbf{b} = -h\mathbf{F}(\mathbf{x})$.

Paso 2 Para $i = 1, 2, \ldots, N$ haga los pasos 3–7.



Paso 3 Determine $A = J(\mathbf{x})$;
Resuelva el sistema lineal $A\mathbf{k}_1 = \mathbf{b}$.

Paso 4 Determine $A = J(\mathbf{x} + \frac{1}{2}\mathbf{k}_1)$;
Resuelva el sistema lineal $A\mathbf{k}_2 = \mathbf{b}$.

Paso 5 Determine $A = J(\mathbf{x} + \frac{1}{2}\mathbf{k}_2)$;
Resuelva el sistema lineal $A\mathbf{k}_3 = \mathbf{b}$.

Paso 6 Determine $A = J(\mathbf{x} + \mathbf{k}_3)$;
Resuelva el sistema lineal $A\mathbf{k}_3 = \mathbf{b}$.

Paso 7 Determine $\mathbf{x} = \mathbf{x} + (\mathbf{k}_1 + 2\mathbf{k}_2 + 2\mathbf{k}_3 + \mathbf{k}_4)/6$.

Paso 8 SALIDA $(x_1, x_2, \ldots, x_n)$;
PARE.



La sección Conjunto de ejercicios 10.5 está disponible en línea. Encuentre la ruta de acceso en las páginas preliminares.



10.6 Software numérico

El paquete Hompack en Netlib resuelve un sistema de ecuaciones no lineales al utilizar varios métodos de homotopía.

Los sistemas no lineales en las bibliotecas IMSL y NAG usan el método Levenberg-Marquardt, que es un promedio ponderado del método de Newton y el método de descenso más rápido. El peso se sesga hacia el método de descenso más rápido hasta que se detecta la convergencia, tiempo en el que el peso cambia hacia un método de Newton que converge más rápido. En cualquier rutina, se puede usar una aproximación de diferencia finita para el jacobiano o una subrutina proporcionada por el usuario para calcular el jacobiano.

Las secciones Preguntas de análisis, Conceptos clave y Revisión del capítulo están disponibles en línea. Encuentre la ruta de acceso en las páginas preliminares.

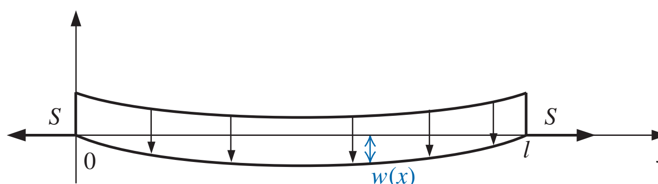
CAPÍTULO

11

Problemas de valor en la frontera para ecuaciones diferenciales ordinarias

Introducción

Un problema común en ingeniería civil concierne a la deflexión de una viga de sección transversal rectangular sujeta a carga uniforme, mientras los extremos de la viga están sujetos de tal forma que no sufren deflexión.



Suponga que l, q, E, S e I representan, respectivamente, la longitud de la viga, la intensidad de la carga uniforme, el módulo de elasticidad, el esfuerzo en los extremos y el momento central de inercia. La ecuación diferencial que aproxima la situación física es de la forma

$$\frac{d^2w}{dx^2}(x) = \frac{S}{EI}w(x) + \frac{qx}{2EI}(x-l),$$

donde $w(x)$ es la deflexión de una distancia x desde el extremo izquierdo de la viga. Puesto que no se presenta deflexión en los extremos de la viga, existen dos condiciones en la frontera:

$$w(0) = 0 \quad \text{y} \quad w(l) = 0.$$

Cuando la viga tiene un grosor uniforme, el producto EI es constante. En este caso, la solución exacta se obtiene fácilmente. Cuando el grosor no es uniforme, el momento de inercia I es una función de x, y se necesitan técnicas de aproximación. Los problemas de este tipo se consideran en el ejercicio 7 de la sección 11.3, en el ejercicio 6 de la sección 11.4 y en el ejercicio 7 de la sección 11.5.

Las ecuaciones diferenciales en el capítulo 5 son de primer orden y tienen una condición inicial que satisfacer. Más adelante en ese capítulo observamos que las técnicas se podrían ampliar a los sistemas de ecuaciones y, después, a ecuaciones de orden superior, pero todas las condiciones específicas están en el mismo extremo. Estos son problemas de valor inicial. En este capítulo mostramos cómo aproximar la solución a los problemas de **valor en la frontera**, ecuaciones diferenciales con condiciones impuestas en diferentes puntos. Para ecuaciones diferenciales de primer orden, sólo se especifica una condición, por lo que no existe distinción entre los problemas de valor inicial y de valor en la frontera. Nosotros consideraremos ecuaciones de segundo orden con dos valores en la frontera.

A menudo, los problemas físicos que dependen de la posición en lugar del tiempo se describen en términos de ecuaciones diferenciales con condiciones impuestas en más de un punto.

En este capítulo, los problemas de valor en la frontera de dos puntos implican una ecuaciones diferencial de segundo orden de la forma

$$y'' = f(x, y, y'), \quad \text{para } a \leq x \leq b, \tag{11.1}$$

junto con las condiciones de frontera

$$y(a) = \alpha \quad \text{y} \quad y(b) = \beta. \tag{11.2}$$

11.1 El método de disparo lineal

El siguiente teorema establece las condiciones generales que garantizan la existencia de la solución para un problema de valor en la frontera de segundo orden. La prueba de este teorema se puede encontrar en [Keller, H].

Teorema 11.1 Suponga que la función f en el problema de valor en la frontera

$$y'' = f(x, y, y'), \quad \text{para } a \leq x \leq b, \text{con } y(a) = \alpha \text{ y } y(b) = \beta,$$

es continua en el conjunto

$$D = \{\, (x, y, y') \mid \text{para } a \leq x \leq b, \text{con } -\infty < y < \infty \text{ y } -\infty < y' < \infty \,\},$$

y que las derivadas parciales f_y y $f_{y'}$ también son continuas en D. Si

i) $f_y(x, y, y') > 0,$ para todas $(x, y, y') \in D,$ y

ii) existe una constante M, con

$$\left| f_{y'}(x, y, y') \right| \leq M, \text{ para todas } (x, y, y') \in D,$$

entonces el problema de valor en la frontera tiene una única solución. ■

Ejemplo 1 Use el teorema 11.1 para mostrar que el problema de valor en la frontera

$$y'' + e^{-xy} + \text{sen}\, y' = 0, \quad \text{para } 1 \leq x \leq 2, \text{con } y(1) = y(2) = 0,$$

tiene una única solución.

Solución tenemos

$$f(x, y, y') = -e^{-xy} - \text{sen}\, y'$$

y, para todas las x en $[1, 2]$,

$$f_y(x, y, y') = xe^{-xy} > 0 \quad \text{y} \quad \left| f_{y'}(x, y, y') \right| = |-\cos y'| \leq 1.$$

Por lo que el problema tiene una única solución. ■

Problema lineal de valor en la frontera

La ecuación diferencial

$$y'' = f(x, y, y')$$

es lineal cuando las funciones $p(x)$, $q(x)$ y $r(x)$ existen con

Una ecuación lineal implica solamente potencias lineales de y y sus derivadas.

$$f(x, y, y') = p(x)y' + q(x)y + r(x).$$

Los problemas de este tipo se presentan con frecuencia y en esta situación el teorema 11.1 se puede simplificar.

Corolario 11.2 Suponga que el problema lineal de valor en la frontera

$$y'' = p(x)y' + q(x)y + r(x), \quad \text{con } a \le x \le b, \text{con } y(a) = \alpha \text{ y } y(b) = \beta,$$

satisface

i) $p(x)$, $q(x)$, y $r(x)$ son continuas en $[a, b]$,

ii) $q(x) > 0$ en $[a, b]$.

Entonces el problema de valor en la frontera tiene una única solución. ■

Para aproximar la solución única de este problema lineal, primero consideramos los problemas de valor inicial

$$y'' = p(x)y' + q(x)y + r(x), \text{ con } \quad a \le x \le b, \quad y(a) = \alpha, \text{ y } \quad y'(a) = 0, \tag{11.3}$$

y

$$y'' = p(x)y' + q(x)y, \text{ con } \quad a \le x \le b, \quad y(a) = 0, \text{ y } \quad y'(a) = 1. \tag{11.4}$$

El teorema 5.17 en la sección 5.9 (consultar la página 248) garantiza que de acuerdo con la hipótesis en el corolario 11.2, ambos problemas tienen solución única.

Sea que $y_1(x)$ denota la solución para la ecuación (11.3) y sea que $y_2(x)$ denota la solución para la ecuación (11.4). Suponga que $y_2(b) \neq 0$. (Que $y_2(b) = 0$ está en conflicto con la hipótesis del corolario 11.2 se considera en el ejercicio 8.) Defina

$$y(x) = y_1(x) + \frac{\beta - y_1(b)}{y_2(b)} y_2(x). \tag{11.5}$$

Entonces $y(x)$ es la solución del problema lineal en la frontera (11.3). Para observarlo, primero note que

$$y'(x) = y_1'(x) + \frac{\beta - y_1(b)}{y_2(b)} y_2'(x)$$

y

$$y''(x) = y_1''(x) + \frac{\beta - y_1(b)}{y_2(b)} y_2''(x).$$

Al sustituir para $y_1''(x)$ y $y_2''(x)$ en esta ecuación obtenemos

$$\begin{aligned} y'' &= p(x)y_1' + q(x)y_1 + r(x) + \frac{\beta - y_1(b)}{y_2(b)} \left(p(x)y_2' + q(x)y_2\right) \\ &= p(x)\left(y_1' + \frac{\beta - y_1(b)}{y_2(b)} y_2'\right) + q(x)\left(y_1 + \frac{\beta - y_1(b)}{y_2(b)} y_2\right) + r(x) \\ &= p(x)y'(x) + q(x)y(x) + r(x). \end{aligned}$$

Además,

$$y(a) = y_1(a) + \frac{\beta - y_1(b)}{y_2(b)} y_2(a) = \alpha + \frac{\beta - y_1(b)}{y_2(b)} \cdot 0 = \alpha$$

y

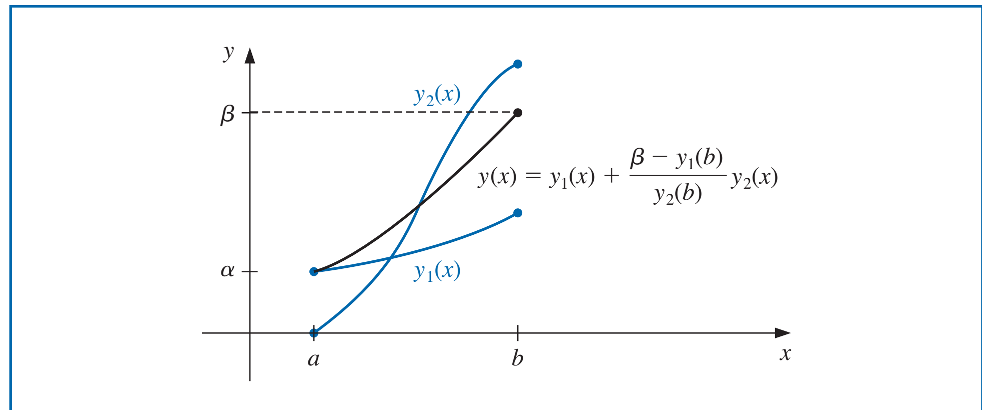
$$y(b) = y_1(b) + \frac{\beta - y_1(b)}{y_2(b)} y_2(b) = y_1(b) + \beta - y_1(b) = \beta.$$

Disparo lineal

Este "disparo" golpea el objetivo después de un disparo de prueba. En la siguiente sección observamos que problemas no lineales requieren múltiples disparos.

El método de disparo para ecuaciones lineales está basado en el reemplazo del problema lineal de valor en la frontera mediante los dos problemas de valor inicial (11.3) y (11.4). Existen numerosos métodos a partir del capítulo 5 para aproximar las soluciones $y_1(x)$ y $y_2(x)$ y una vez que estas aproximaciones están disponibles, la solución para el problema de valor en la frontera se aproxima usando la ecuación (11.5). Gráficamente, el método tiene el aspecto mostrado en la figura 11.1

Figura 11.1



El algoritmo 11.1 usa la técnica Runge-Kutta de cuarto orden para encontrar la aproximación para $y_1(x)$ y $y_2(x)$, pero otras técnicas para aproximar las soluciones para problemas de valor inicial se pueden sustituir en el paso 4.

Primero, escribimos la ecuación (11.3) como un sistema de dos ecuaciones diferenciales al permitir $z_1(x) = y(x)$ y $z_2(x) = y'(x)$ de tal forma que

$$z_1'(x) = z_2(x)$$
$$z_2'(x) = p(x)z_2(x) + q(x)z_1(x) + r(x)$$

para $a \le x \le b$ con $z_1(a) = \alpha$ y $z_2(a) = 0$. A continuación, escribimos la ecuación (11.4) como un sistema de dos ecuaciones diferenciales lineales haciendo $z_3(x) = y(x)$ y $z_4(x) = y'(x)$ de tal forma que

$$z_3'(x) = z_4(x)$$
$$z_4'(x) = p(x)z_4(x) + q(x)z_3(x)$$

para $a \le x \le b$ con $z_3(a) = 0$ y $z_4(a) = 1$. Las aproximaciones calculadas en el algoritmo son

$$u_{1,i} \approx z_1(x_i) = y_1(x_i), \quad u_{2,i} \approx z_2(x_i) = y_1'(x_i)$$

y

$$v_{1,i} \approx z_3(x_i) = y_2(x_i), \quad v_{2,i} \approx z_4(x_i) = y_2'(x_i).$$

Las aproximaciones finales son

$$w_{1,i} = u_{1,i} + \frac{\beta - u_{1,N}}{v_{1,N}} v_{1,i} \approx y_1(x_i)$$

y

$$w_{2,i} = u_{2,i} + \frac{\beta - u_{1,N}}{v_{1,N}} v_{2,i} \approx y_1'(x_i)$$

El algoritmo tiene la característica adicional de obtener aproximaciones para la derivada de la solución del problema de valor en la frontera, así como para la solución del problema mismo. El uso del algoritmo no está restringido a los problemas para los que las hipótesis del corolario 11.2 se pueden verificar; sería útil para muchos problemas que no satisfacen estas hipótesis. Un ejemplo de este tipo se puede encontrar en el ejercicio 4.

ALGORITMO

11.1

Disparo lineal

Para aproximar la solución del problema de valor en la frontera

$$-y'' + p(x)y' + q(x)y + r(x) = 0, \quad \text{para } a \leq x \leq b, \text{ con } y(a) = \alpha \text{ y } y(b) = \beta,$$

(*Nota*: *Las ecuaciones* (11.3) *y* (11.4) *se escriben como sistemas de primer orden y se resuelven*.)

ENTRADA extremos a, b; condiciones de frontera α, β; número de subintervalos N.

SALIDA aproximaciones $w_{1,i}$ para $y(x_i)$; $w_{2,i}$ para $y'(x_i)$ por cada $i = 0, 1, \ldots, N$.

Paso 1 Determine $h = (b - a)/N$;
$u_{1,0} = \alpha$;
$u_{2,0} = 0$;
$v_{1,0} = 0$;
$v_{2,0} = 1$.

Paso 2 Para $i = 0, \ldots, N - 1$ haga los pasos 3 y 4.
(*Se usa el método Runge-Kutta para sistemas en los pasos* 3 *y* 4.)

Paso 3 Determine $x = a + ih$.

Paso 4 Determine $k_{1,1} = hu_{2,i}$;

$$k_{1,2} = h\left[p(x)u_{2,i} + q(x)u_{1,i} + r(x)\right];$$
$$k_{2,1} = h\left[u_{2,i} + \tfrac{1}{2}k_{1,2}\right];$$
$$k_{2,2} = h\left[p(x + h/2)\left(u_{2,i} + \tfrac{1}{2}k_{1,2}\right) + q(x + h/2)\left(u_{1,i} + \tfrac{1}{2}k_{1,1}\right) + r(x + h/2)\right];$$
$$k_{3,1} = h\left[u_{2,i} + \tfrac{1}{2}k_{2,2}\right];$$
$$k_{3,2} = h\left[p(x + h/2)\left(u_{2,i} + \tfrac{1}{2}k_{2,2}\right) + q(x + h/2)(u_{1,i} + \tfrac{1}{2}k_{2,1}) + r(x + h/2)\right];$$
$$k_{4,1} = h\left[u_{2,i} + k_{3,2}\right];$$
$$k_{4,2} = h\left[p(x + h)(u_{2,i} + k_{3,2}) + q(x + h)(u_{1,i} + k_{3,1}) + r(x + h)\right];$$

$$u_{1,i+1} = u_{1,i} + \tfrac{1}{6}\left[k_{1,1} + 2k_{2,1} + 2k_{3,1} + k_{4,1}\right];$$
$$u_{2,i+1} = u_{2,i} + \tfrac{1}{6}\left[k_{1,2} + 2k_{2,2} + 2k_{3,2} + k_{4,2}\right];$$
$$k'_{1,1} = hv_{2,i};$$
$$k'_{1,2} = h\left[p(x)v_{2,i} + q(x)v_{1,i}\right];$$
$$k'_{2,1} = h\left[v_{2,i} + \tfrac{1}{2}k'_{1,2}\right];$$
$$k'_{2,2} = h\left[p(x+h/2)\left(v_{2,i} + \tfrac{1}{2}k'_{1,2}\right) + q(x+h/2)\left(v_{1,i} + \tfrac{1}{2}k'_{1,1}\right)\right];$$
$$k'_{3,1} = h\left[v_{2,i} + \tfrac{1}{2}k'_{2,2}\right];$$
$$k'_{3,2} = h\left[p(x+h/2)\left(v_{2,i} + \tfrac{1}{2}k'_{2,2}\right) + q(x+h/2)\left(v_{1,i} + \tfrac{1}{2}k'_{2,1}\right)\right];$$
$$k'_{4,1} = h\left[v_{2,i} + k'_{3,2}\right];$$
$$k'_{4,2} = h\left[p(x+h)(v_{2,i} + k'_{3,2}) + q(x+h)(v_{1,i} + k'_{3,1})\right];$$
$$v_{1,i+1} = v_{1,i} + \tfrac{1}{6}\left[k'_{1,1} + 2k'_{2,1} + 2k'_{3,1} + k'_{4,1}\right];$$
$$v_{2,i+1} = v_{2,i} + \tfrac{1}{6}\left[k'_{1,2} + 2k'_{2,2} + 2k'_{3,2} + k'_{4,2}\right].$$

Paso 5 Determine $w_{1,0} = \alpha;$

$$w_{2,0} = \frac{\beta - u_{1,N}}{v_{1,N}};$$

SALIDA $(a, w_{1,0}, w_{2,0})$.

Paso 6 Para $i = 1, \ldots, N$

determine $W1 = u_{1,i} + w_{2,0}v_{1,i};$

$W2 = u_{2,i} + w_{2,0}v_{2,i};$

$x = a + ih;$

SALIDA $(x, W1, W2)$. (*La salida es* $x_i, w_{1,i}, w_{2,i}$.)

Paso 7 PARE. (*El proceso está completo*.) ■

Ejemplo 2 Aplique la técnica de disparo lineal con $N = 10$ al problema de valor en la frontera

$$y'' = -\frac{2}{x}y' + \frac{2}{x^2}y + \frac{\operatorname{sen}(\ln x)}{x^2}, \quad \text{para } 1 \le x \le 2, \text{ con } y(1) = 1 \text{ y } y(2) = 2,$$

y compare los resultados con los de la solución exacta

$$y = c_1 x + \frac{c_2}{x^2} - \frac{3}{10}\operatorname{sen}(\ln x) - \frac{1}{10}\cos(\ln x),$$

donde

$$c_2 = \frac{1}{70}[8 - 12\operatorname{sen}(\ln 2) - 4\cos(\ln 2)] \approx -0.03920701320$$

y

$$c_1 = \frac{11}{10} - c_2 \approx 1.1392070132.$$

Solución Aplicar el algoritmo 11.1 a este problema requiere aproximar las soluciones para los problemas de valor inicial

$$y_1'' = -\frac{2}{x}y_1' + \frac{2}{x^2}y_1 + \frac{\operatorname{sen}(\ln x)}{x^2}, \quad \text{para } 1 \le x \le 2, \text{ con } y_1(1) = 1 \text{ y } y_1'(1) = 0,$$

Si esta técnica de disparo inverso sigue dando la cancelación de los dígitos significativos y si el incremento de precisión no arroja mayor exactitud, es preciso usar otras técnicas. Algunas de ellas se presentan más adelante en este capítulo. Sin embargo, en general, si $u_{1,i}$ y $v_{1,i}$ son $O(h^n)$ aproximaciones para $y_1(x_i)$ y $y_2(x_i)$, respectivamente, para cada $i = 0, 1, \ldots, N$, entonces $w_{1,i}$ será una aproximación $O(h^n)$ para $y(x_i)$. En particular,

$$|w_{1,i} - y(x_i)| \leq Kh^n \left| 1 + \frac{v_{1,i}}{v_{1,N}} \right|,$$

para alguna constante K (consulte [IK], p. 426).

La sección Conjunto de ejercicios 11.1 está disponible en línea. Encuentre la ruta de acceso en las páginas preliminares.

11.2 El método de disparo para problemas no lineales

La técnica de disparo para el problema no lineal de valor en la frontera de segundo orden

$$y'' = f(x, y, y'), \quad \text{para } a \leq x \leq b, \text{ con } y(a) = \alpha \text{ y } y(b) = \beta, \tag{11.6}$$

es similar a la técnica lineal, excepto que la solución para un problema no lineal no se puede expresar como una combinación lineal de soluciones para dos problemas de valor inicial. Por el contrario, aproximamos la solución al problema de valor en la frontera por medio de las soluciones para una *sucesión* de problemas de valor inicial que implican un parámetro t. Estos problemas tienen la forma

$$y'' = f(x, y, y'), \quad \text{para } a \leq x \leq b, \text{ con } y(a) = \alpha \text{ y } y'(a) = t. \tag{11.7}$$

Lo hacemos al seleccionar los parámetros $t = t_k$ de forma que se garantiza

$$\lim_{k \to \infty} y(b, t_k) = y(b) = \beta,$$

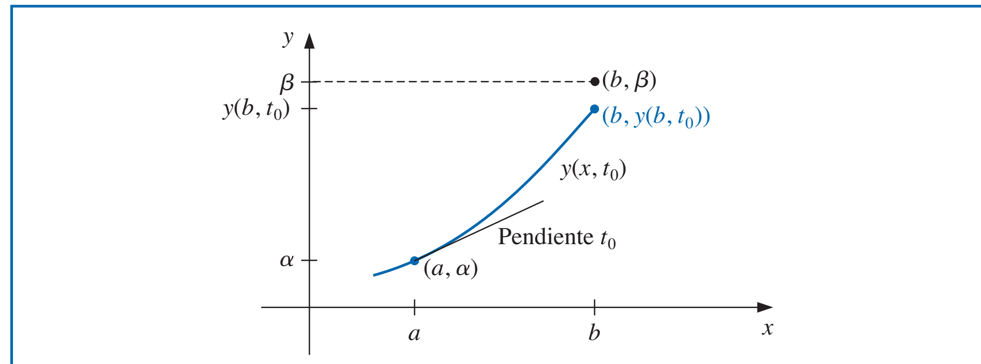
Los métodos de disparo para los problemas no lineales requieren iteraciones para acercarse al "objetivo".

donde $y(x, t_k)$ denota la solución del problema de valor inicial (11.7) con $t = t_k$ y $y(x)$ denota la solución del problema de valor en la frontera (11.6).

Esta técnica recibe el nombre de método de "disparo" por la analogía con el procedimiento de disparar a objetos en un objetivo inmóvil. (Consulte la figura 11.2.) Comenzamos con un parámetro t_0 que determina la elevación inicial a la que se dispara al objeto desde el punto (a, α) y a lo largo de la curva descrita por medio de la solución del problema de valor inicial:

$$y'' = f(x, y, y'), \quad \text{para } a \leq x \leq b, \text{ con } y(a) = \alpha \text{ y } y'(a) = t_0.$$

Figura 11.2



Si $y(b, t_0)$ no está suficientemente cerca de β, corregimos nuestra aproximación al seleccionar elevaciones t_1, t_2, y así sucesivamente, hasta que $y(b, t_k)$ esté suficientemente cerca de "golpear" β. (Consulte la figura 11.3.)

Para determinar los parámetros t_k, suponga que un problema de valor en la frontera de la forma (11.6) satisface las hipótesis del teorema 11.1. Si $y(x, t)$ denota la solución del problema de valor inicial (11.7), a continuación determinamos t con

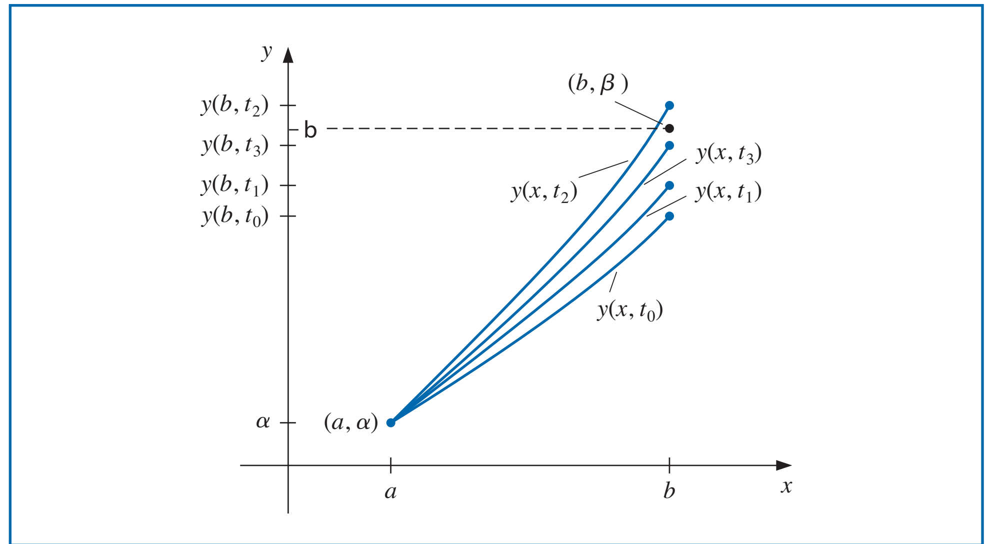
$$y(b, t) - \beta = 0. \tag{11.8}$$

Ésta es una ecuación no lineal en la variable t. Los problemas de este tipo se consideraron en el capítulo 2 y existen diferentes métodos.

Para utilizar el método de la secante para resolver el problema, necesitamos seleccionar aproximaciones iniciales t_0 y t_1, y, después, generar los términos restantes de la sucesión a través de

$$t_k = t_{k-1} - \frac{(y(b, t_{k-1}) - \beta)(t_{k-1} - t_{k-2})}{y(b, t_{k-1}) - y(b, t_{k-2})}, \quad k = 2, 3, \ldots .$$

Figura 11.3



Iteración de Newton

Para usar el método de Newton más poderoso para generar la sucesión $\{t_k\}$, sólo se necesita una aproximación inicial t_0. Sin embargo, la iteración tiene la forma

$$t_k = t_{k-1} - \frac{y(b, t_{k-1}) - \beta}{\frac{dy}{dt}(b, t_{k-1})}, \tag{11.9}$$

y requiere el conocimiento de $(dy/dt)(b, t_{k-1})$. Esto presenta una dificultad porque no se conoce una representación explícita para $y(b, t)$; solamente conocemos los valores $y(b, t_0), y(b, t_1), \ldots, y(b, t_{k-1})$.

Suponga que reescribimos el problema de valor inicial (11.7), al enfatizar que la solución depende tanto de x como del parámetro t:

$$y''(x, t) = f(x, y(x, t), y'(x, t)), \text{ para } a \leq x \leq b, \text{con } y(a, t) = \alpha \text{ y } y'(a, t) = t. \tag{11.10}$$

Paso 1 Determine $h = (b - a)/N$;
$k = 1$;
$TK = (\beta - \alpha)/(b - a)$. (*Nota: TK también podría ser una entrada.*)

Paso 2 Mientras ($k \leq M$) haga los pasos 3–10.

Paso 3 Determine $w_{1,0} = \alpha$;
$w_{2,0} = TK$;
$u_1 = 0$;
$u_2 = 1$.

Paso 4 Para $i = 1, \ldots, N$ haga los pasos 5 y 6.
(*Se utiliza el método Runge-Kutta para sistemas en los pasos* 5 *y* 6.)

Paso 5 Determine $x = a + (i - 1)h$.

Paso 6 Determine $k_{1,1} = hw_{2,i-1}$;

$$k_{1,2} = hf(x, w_{1,i-1}, w_{2,i-1});$$
$$k_{2,1} = h\left(w_{2,i-1} + \tfrac{1}{2}k_{1,2}\right);$$
$$k_{2,2} = hf\left(x + h/2, w_{1,i-1} + \tfrac{1}{2}k_{1,1}, w_{2,i-1} + \tfrac{1}{2}k_{1,2}\right);$$
$$k_{3,1} = h\left(w_{2,i-1} + \tfrac{1}{2}k_{2,2}\right);$$
$$k_{3,2} = hf\left(x + h/2, w_{1,i-1} + \tfrac{1}{2}k_{2,1}, w_{2,i-1} + \tfrac{1}{2}k_{2,2}\right);$$
$$k_{4,1} = h(w_{2,i-1} + k_{3,2});$$
$$k_{4,2} = hf(x + h, w_{1,i-1} + k_{3,1}, w_{2,i-1} + k_{3,2});$$
$$w_{1,i} = w_{1,i-1} + (k_{1,1} + 2k_{2,1} + 2k_{3,1} + k_{4,1})/6;$$
$$w_{2,i} = w_{2,i-1} + (k_{1,2} + 2k_{2,2} + 2k_{3,2} + k_{4,2})/6;$$
$$k'_{1,1} = hu_2;$$
$$k'_{1,2} = h[f_y(x, w_{1,i-1}, w_{2,i-1})u_1 + f_{y'}(x, w_{1,i-1}, w_{2,i-1})u_2];$$
$$k'_{2,1} = h\left[u_2 + \tfrac{1}{2}k'_{1,2}\right];$$
$$k'_{2,2} = h\left[f_y(x + h/2, w_{1,i-1}, w_{2,i-1})\left(u_1 + \tfrac{1}{2}k'_{1,1}\right) + f_{y'}(x + h/2, w_{1,i-1}, w_{2,i-1})\left(u_2 + \tfrac{1}{2}k'_{1,2}\right)\right];$$
$$k'_{3,1} = h\left(u_2 + \tfrac{1}{2}k'_{2,2}\right);$$
$$k'_{3,2} = h\left[f_y(x + h/2, w_{1,i-1}, w_{2,i-1})\left(u_1 + \tfrac{1}{2}k'_{2,1}\right) + f_{y'}(x + h/2, w_{1,i-1}, w_{2,i-1})\left(u_2 + \tfrac{1}{2}k'_{2,2}\right)\right];$$
$$k'_{4,1} = h(u_2 + k'_{3,2});$$
$$k'_{4,2} = h\left[f_y(x + h, w_{1,i-1}, w_{2,i-1})\left(u_1 + k'_{3,1}\right) + f_{y'}(x + h, w_{1,i-1}, w_{2,i-1})\left(u_2 + k'_{3,2}\right)\right];$$
$$u_1 = u_1 + \tfrac{1}{6}[k'_{1,1} + 2k'_{2,1} + 2k'_{3,1} + k'_{4,1}];$$
$$u_2 = u_2 + \tfrac{1}{6}[k'_{1,2} + 2k'_{2,2} + 2k'_{3,2} + k'_{4,2}].$$

Paso 7 Si $|w_{1,N} - \beta| \leq TOL$ entonces haga los pasos 8 y 9.

Paso 8 Para $i = 0, 1, \ldots, N$
determine $x = a + ih$;
SALIDA $(x, w_{1,i}, w_{2,i})$.

Paso 9 (*El procedimiento está completo*.)
PARE.

Paso 0 Determine $TK = TK - \dfrac{w_{1,N} - \beta}{u_1}$;

(*Se utiliza el método de Newton para calcular* TK.)

$k = k + 1$.

Paso 11 SALIDA ('Número máximo de iteraciones excedido');

(*El procedimiento no fue exitoso* .)

PARE. ■

El valor $t_0 = TK$ seleccionado en el paso 1 es la pendiente de la recta que pasa por (a, α) y (b, β). Si el problema satisface la hipótesis del teorema 11.1, cualquier selección de t_0 proporcionará convergencia, pero una buena selección de t_0 mejorará la convergencia y el procedimiento funcionará para muchos problemas que no satisfacen esta hipótesis. Un ejemplo de este tipo se puede encontrar en el ejercicio 3d).

Ejemplo 1 Aplique el método de disparo con el método de Newton al problema de valor en la frontera

$$y'' = \frac{1}{8}(32 + 2x^3 - yy'), \quad \text{para } 1 \le x \le 3, \text{con } y(1) = 17 \text{ y } y(3) = \frac{43}{3}.$$

Use $N = 20$, $M = 10$, y $TOL = 10^{-5}$ y compare los resultados con la solución exacta $y(x) = x^2 + 16/x$.

Solución Necesitamos aproximar las soluciones de los problemas de valor inicial

$$y'' = \frac{1}{8}(32 + 2x^3 - yy'), \quad \text{para } 1 \le x \le 3, \text{con } y(1) = 17 \text{ y } y'(1) = t_k,$$

y

$$z'' = \frac{\partial f}{\partial y} z + \frac{\partial f}{\partial y'} z' = -\frac{1}{8}(y'z + yz'), \quad \text{para } 1 \le x \le 3, \text{con } z(1) = 0 \text{ y } z'(1) = 1,$$ ■

en cada paso en la iteración. Si la técnica de paro en el algoritmo 11.2 requiere

$$|w_{1,N}(t_k) - y(3)| \le 10^{-5},$$

entonces necesitamos cuatro iteraciones y $t_4 = -14.000203$. Los resultados obtenidos para este valor de t se muestran en la tabla 11.2.

Tabla 11.2

x_i	$w_{1,i}$	$y(x_i)$	$\lvert w_{1,i} - y(x_i)\rvert$
1.0	17.000000	17.000000	
1.1	15.755495	15.755455	4.06×10^{-5}
1.2	14.773389	14.773333	5.60×10^{-5}
1.3	13.997752	13.997692	5.94×10^{-5}
1.4	13.388629	13.388571	5.71×10^{-5}
1.5	12.916719	12.916667	5.23×10^{-5}
1.6	12.560046	12.560000	4.64×10^{-5}
1.7	12.301805	12.301765	4.02×10^{-5}
1.8	12.128923	12.128889	3.14×10^{-5}
1.9	12.031081	12.031053	2.84×10^{-5}
2.0	12.000023	12.000000	2.32×10^{-5}
2.1	12.029066	12.029048	1.84×10^{-5}
2.2	12.112741	12.112727	1.40×10^{-5}
2.3	12.246532	12.246522	1.01×10^{-5}
2.4	12.426673	12.426667	6.68×10^{-6}
2.5	12.650004	12.650000	3.61×10^{-6}
2.6	12.913847	12.913845	9.17×10^{-7}
2.7	13.215924	13.215926	1.43×10^{-6}
2.8	13.554282	13.554286	3.46×10^{-6}
2.9	13.927236	13.927241	5.21×10^{-6}
3.0	14.333327	14.333333	6.69×10^{-6}

A pesar de que el método de Newton que se usó con la técnica de disparo requiere la solución de un problema de valor inicial adicional, por lo general dará convergencia más rápida que el método de secante. Sin embargo, ambos métodos sólo son localmente convergentes porque requieren buenas aproximaciones iniciales.

Para un análisis general de la convergencia de las técnicas de disparo para los problemas no lineales, se refiere al lector al excelente libro de Keller [Keller, H.] En esa referencia se analizan condiciones de frontera más generales. También se debe observar que la técnica de disparo para los problemas no lineales es sensible a los errores de redondeo, en especial si las soluciones $y(x)$ y $z(x, t)$ son funciones de x que crecen rápidamente en $[a, b]$.

La sección Conjunto de ejercicios 11.2 está disponible en línea. Encuentre la ruta de acceso en las páginas preliminares.

11.3 Métodos de diferencias finitas para problemas lineales

Los métodos de disparo lineal y no lineal para los problemas de valor en la frontera pueden presentar problemas de inestabilidad. Los métodos en esta sección tienen mejores características de estabilidad, pero, en general, requieren más cálculos para obtener una exactitud específica.

Los métodos que implican diferencias finitas para resolver los problemas de valor en la frontera reemplazan cada una de las derivadas en la ecuación diferencial con una aproximación de cociente de diferencia adecuada del tipo considerado en la sección 4.1. El cociente de diferencia particular y la longitud de paso h se seleccionan para mantener un orden específico de error de truncamiento. Sin embargo, h no se puede elegir demasiado pequeña debido a la inestabilidad general de las aproximaciones de la derivada.

Aproximación discreta

El método de diferencias finitas para el problema lineal de valor en la frontera de segundo orden,

$$y'' = p(x)y' + q(x)y + r(x), \text{ para } a \leq x \leq b, \text{ con } y(a) = \alpha \text{ y } y(b) = \beta, \quad (11.14)$$

requiere que se usen aproximaciones de cociente de diferencia para aproximar tanto y' como y''. Primero, seleccionamos un entero $N > 0$ y dividimos el intervalo $[a, b]$ en $(N + 1)$ subintervalos iguales, cuyos extremos son los puntos de malla $x_i = a + ih$, para $i = 0, 1, \ldots, N + 1$, donde $h = (b - a)/(N + 1)$. Seleccionar un tamaño de longitud h de esta forma facilita la aplicación de un algoritmo matricial del capítulo 6, lo cual resuelve un sistema lineal que implica una matriz $N \times N$.

En los puntos de malla interior, x_i, para $i = 1, 2, \ldots, N$, la ecuación diferencial a aproximar es

$$y''(x_i) = p(x_i)y'(x_i) + q(x_i)y(x_i) + r(x_i). \quad (11.15)$$

Al expandir y en el tercer polinomio de Taylor alrededor de x_i evaluado en x_{i+1} y x_{i-1}, tenemos, suponiendo que $y \in C^4[x_{i-1}, x_{i+1}]$,

$$y(x_{i+1}) = y(x_i + h) = y(x_i) + hy'(x_i) + \frac{h^2}{2}y''(x_i) + \frac{h^3}{6}y'''(x_i) + \frac{h^4}{24}y^{(4)}(\xi_i^+),$$

para alguna ξ_i^+ en (x_i, x_{i+1}), y

$$y(x_{i-1}) = y(x_i - h) = y(x_i) - hy'(x_i) + \frac{h^2}{2}y''(x_i) - \frac{h^3}{6}y'''(x_i) + \frac{h^4}{24}y^{(4)}(\xi_i^-),$$

para alguna ξ_i^- en (x_{i-1}, x_i). Si se suman estas ecuaciones, tenemos

$$y(x_{i+1}) + y(x_{i-1}) = 2y(x_i) + h^2 y''(x_i) + \frac{h^4}{24}[y^{(4)}(\xi_i^+) + y^{(4)}(\xi_i^-)],$$

y al resolver para $y''(x_i)$ obtenemos

$$y''(x_i) = \frac{1}{h^2}[y(x_{i+1}) - 2y(x_i) + y(x_{i-1})] - \frac{h^2}{24}[y^{(4)}(\xi_i^+) + y^{(4)}(\xi_i^-)].$$

Podemos usar el teorema de valor intermedio 1.11 para simplificar el término de error para proporcionar

$$y''(x_i) = \frac{1}{h^2}[y(x_{i+1}) - 2y(x_i) + y(x_{i-1})] - \frac{h^2}{12} y^{(4)}(\xi_i), \tag{11.16}$$

para alguna ξ_i en (x_{i-1}, x_{i+1}). Esto recibe el nombre de **fórmula de diferencia centrada** para $y''(x_i)$.

Una fórmula de diferencia centrada para $y'(x_i)$ se obtiene de forma similar (los detalles se consideraron en la sección 4.1), lo cual resulta en

$$y'(x_i) = \frac{1}{2h}[y(x_{i+1}) - y(x_{i-1})] - \frac{h^2}{6} y'''(\eta_i), \tag{11.17}$$

para alguna η_i en (x_{i-1}, x_{i+1}).

El uso de estas fórmulas de diferencia centrada en la ecuación (11.15) genera la ecuación

$$\frac{y(x_{i+1}) - 2y(x_i) + y(x_{i-1})}{h^2} = p(x_i)\left[\frac{y(x_{i+1}) - y(x_{i-1})}{2h}\right] + q(x_i)y(x_i)$$
$$+ r(x_i) - \frac{h^2}{12}\left[2p(x_i)y'''(\eta_i) - y^{(4)}(\xi_i)\right].$$

Un método de diferencias finitas con error de truncamiento de orden $O(h^2)$ resulta del uso de esta ecuación con las condiciones de frontera $y(a) = \alpha$ y $y(b) = \beta$ para definir el sistema de ecuaciones lineales

$$w_0 = \alpha, \qquad w_{N+1} = \beta$$

y

$$\left(\frac{-w_{i+1} + 2w_i - w_{i-1}}{h^2}\right) + p(x_i)\left(\frac{w_{i+1} - w_{i-1}}{2h}\right) + q(x_i)w_i = -r(x_i), \tag{11.18}$$

para cada $i = 1, 2, \ldots, N$.

En la forma que consideraremos, la ecuación (11.18) se reescribe como

$$-\left(1 + \frac{h}{2}p(x_i)\right)w_{i-1} + \left(2 + h^2 q(x_i)\right)w_i - \left(1 - \frac{h}{2}p(x_i)\right)w_{i+1} = -h^2 r(x_i),$$

y el sistema de ecuaciones resultante se expresa en forma de la matriz tridiagonal $N \times N$

$$A\mathbf{w} = \mathbf{b}, \quad \text{donde} \tag{11.19}$$

$$A = \begin{bmatrix} 2+h^2q(x_1) & -1+\frac{h}{2}p(x_1) & 0 & \cdots & \cdots & 0 \\ -1-\frac{h}{2}p(x_2) & 2+h^2q(x_2) & -1+\frac{h}{2}p(x_2) & \ddots & & \vdots \\ 0 & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & \ddots & \ddots & -1+\frac{h}{2}p(x_{N-1}) \\ 0 & \cdots & \cdots & 0 & -1-\frac{h}{2}p(x_N) & 2+h^2q(x_N) \end{bmatrix},$$

$$\mathbf{w} = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_{N-1} \\ w_N \end{bmatrix}, \quad \text{y} \qquad \mathbf{b} = \begin{bmatrix} -h^2r(x_1) + \left(1+\frac{h}{2}p(x_1)\right)w_0 \\ -h^2r(x_2) \\ \vdots \\ -h^2r(x_{N-1}) \\ -h^2r(x_N) + \left(1-\frac{h}{2}p(x_N)\right)w_{N+1} \end{bmatrix}.$$

El siguiente teorema establece las condiciones en las cuales el sistema lineal tridiagonal (11.19) tiene una única solución. Su demostración es una consecuencia del teorema 6.31 en la página 318 y se considera en el ejercicio 9.

Teorema 11.3 Suponga que p, q y r son continuas en $[a, b]$. Si $q(x) \geq 0$ en $[a, b]$, entonces el sistema lineal tridiagonal (11.9) tiene una única solución siempre que $h < 2/L$, donde $L = \text{máx}_{a \leq x \leq b} |p(x)|$. ■

Se debería observar que la hipótesis del teorema 11.3 garantiza una solución única para el problema de valor en la frontera (11.4), pero no garantiza que $y \in C^4[a, b]$. Necesitamos establecer que $y^{(4)}$ es continua en $[a, b]$ para garantizar que el error de truncamiento tiene orden $O(h^2)$.

El algoritmo 11.3 implementa el método de diferencias finitas lineal.

ALGORITMO
11.3

Diferencia finita lineal

Para aproximar la solución del problema de valor en la frontera

$$y'' = p(x)y' + q(x)y + r(x), \quad \text{para } a \leq x \leq b, \text{con } y(a) = \alpha \text{ y } y(b) = \beta:$$

ENTRADA extremos a, b; condiciones de frontera α, β; entero $N \geq 2$.

SALIDA aproximaciones w_i para $y(x_i)$ para cada $i = 0, 1, \ldots, N+1$.

Paso 1 Determine $h = (b-a)/(N+1)$;
$x = a + h$;
$a_1 = 2 + h^2q(x)$;
$b_1 = -1 + (h/2)p(x)$;
$d_1 = -h^2r(x) + (1 + (h/2)p(x))\alpha$.

Paso 2 Para $i = 2, \ldots, N-1$
determine $x = a + ih$;
$a_i = 2 + h^2q(x)$;
$b_i = -1 + (h/2)p(x)$;
$c_i = -1 - (h/2)p(x)$;
$d_i = -h^2r(x)$.

Paso 3 Determine $x = b - h$;
$a_N = 2 + h^2 q(x)$;
$c_N = -1 - (h/2)p(x)$;
$d_N = -h^2 r(x) + (1 - (h/2)p(x))\beta$.

Paso 4 Determine $l_1 = a_1$; (*Los pasos* 4-8 *resuelven un sistema lineal tridiagonal usando el algoritmo* 6.7.)
$u_1 = b_1/a_1$;
$z_1 = d_1/l_1$.

Paso 5 Para $i = 2, \ldots, N-1$ determine $l_i = a_i - c_i u_{i-1}$;
$u_i = b_i/l_i$;
$z_i = (d_i - c_i z_{i-1})/l_i$.

Paso 6 Determine $l_N = a_N - c_N u_{N-1}$;
$z_N = (d_N - c_N z_{N-1})/l_N$.

Paso 7 Determine $w_0 = \alpha$;
$w_{N+1} = \beta$.
$w_N = z_N$.

Paso 8 Para $i = N-1, \ldots, 1$ determine $w_i = z_i - u_i w_{i+1}$.

Paso 9 Para $i = 0, \ldots, N+1$ determine $x = a + ih$;
SALIDA (x, w_i).

Paso 10 PARE. (*El procedimiento está completo*.) ■

Ejemplo 1 Use el algoritmo 11.3 con $N = 9$ para aproximar la solución del problema lineal de valor en la frontera

$$y'' = -\frac{2}{x}y' + \frac{2}{x^2}y + \frac{\operatorname{sen}(\ln x)}{x^2}, \quad \text{para } 1 \le x \le 2, \text{con } y(1) = 1 \text{ y } y(2) = 2,$$

y compare los resultados con los obtenidos con el método de disparo en el ejemplo 2 de la sección 11.1.

Solución Para este ejemplo usaremos $N = 9$ por lo que $h = 0.1$ y tenemos el mismo espaciado que en el ejemplo 2 de la sección 11.1. Los resultados completos se enumeran en la tabla 11.3.

Tabla 11.3

x_i	w_i	$y(x_i)$	$\|w_i - y(x_i)\|$
1.0	1.00000000	1.00000000	
1.1	1.09260052	1.09262930	2.88×10^{-5}
1.2	1.18704313	1.18708484	4.17×10^{-5}
1.3	1.28333687	1.28338236	4.55×10^{-5}
1.4	1.38140205	1.38144595	4.39×10^{-5}
1.5	1.48112026	1.48115942	3.92×10^{-5}
1.6	1.58235990	1.58239246	3.26×10^{-5}
1.7	1.68498902	1.68501396	2.49×10^{-5}
1.8	1.78888175	1.78889853	1.68×10^{-5}
1.9	1.89392110	1.89392951	8.41×10^{-6}
2.0	2.00000000	2.00000000	

Estos resultados son considerablemente menos exactos que los obtenidos en el ejemplo 2 de la sección 11.1. Esto se debe a que el método que se usó en ese ejemplo implicaba una técnica Runge-Kutta con error de truncamiento local de orden $O(h^4)$, mientras el método de diferencia que se usa aquí tiene un error de truncamiento local de orden $O(h^2)$. ■

Para obtener un método de diferencia con mayor exactitud, podemos proceder de diferentes formas. Usando la serie de Taylor de quinto orden para aproximar $y''(x_i)$ y $y'(x_i)$ en un término de error de truncamiento que implica h^4. Sin embargo, este proceso requiere usar múltiplos no sólo de $y(x_{i+1})$ y $y(x_{i-1})$, sino también de $y(x_{i+2})$ y $y(x_{i-2})$ en las fórmulas de aproximación para $y''(x_i)$ y $y'(x_i)$. Esto conduce a la dificultad en $i = N$ porque no conocemos w_{-1} y en $i = N$ porque no conocemos w_{N+2}. Además, el sistema de ecuaciones análogas resultante para (11.19) no es de forma tridiagonal y la solución del sistema requiere muchos más cálculos.

Uso de la extrapolación de Richardson

En lugar de intentar obtener un método de diferencia con un error de truncamiento de orden superior de esta forma, en general, es más satisfactorio considerar una reducción del tamaño de longitud. Además, la técnica de extrapolación de Richardson se puede usar de manera efectiva para este método porque el término de error está expresado en potencias pares de h con coeficientes independientes de h, siempre y cuando y sea suficientemente diferenciable (consulte, por ejemplo, [Keller, H], p. 81).

El ejercicio 10 da algunas ideas respecto a la forma del error de truncamiento y la justificación para usar la extrapolación.

Ejemplo 2 Aplique la extrapolación de Richardson para aproximar la solución del problema de valor en la frontera

$$y'' = -\frac{2}{x}y' + \frac{2}{x^2}y + \frac{\operatorname{sen}(\ln x)}{x^2}, \quad \text{para } 1 \le x \le 2, \text{con } y(1) = 1 \text{ y } y(2) = 2,$$

usando $h = 0.1, 0.05$, y 0.025.

Solución Los resultados se muestran en la tabla 11.4. La primera extrapolación es

$$\text{Ext}_{1i} = \frac{4w_i(h = 0.05) - w_i(h = 0.1)}{3},$$

la segunda extrapolación es

$$\text{Ext}_{2i} = \frac{4w_i(h = 0.025) - w_i(h = 0.05)}{3},$$

y la extrapolación final es

$$\text{Ext}_{3i} = \frac{16\text{Ext}_{2i} - \text{Ext}_{1i}}{15}.$$

Tabla 11.4

x_i	$w_i(h = 0.05)$	$w_i(h = 0.025)$	Ext_{1i}	Ext_{2i}	Ext_{3i}
1.0	1.00000000	1.00000000	1.00000000	1.00000000	1.00000000
1.1	1.09262207	1.09262749	1.09262925	1.09262930	1.09262930
1.2	1.18707436	1.18708222	1.18708477	1.18708484	1.18708484
1.3	1.28337094	1.28337950	1.28338230	1.28338236	1.28338236
1.4	1.38143493	1.38144319	1.38144589	1.38144595	1.38144595
1.5	1.48114959	1.48115696	1.48115937	1.48115941	1.48115942
1.6	1.58238429	1.58239042	1.58239242	1.58239246	1.58239246
1.7	1.68500770	1.68501240	1.68501393	1.68501396	1.68501396
1.8	1.78889432	1.78889748	1.78889852	1.78889853	1.78889853
1.9	1.89392740	1.89392898	1.89392950	1.89392951	1.89392951
2.0	2.00000000	2.00000000	2.00000000	2.00000000	2.00000000

Se omiten los valores de w_i ($h = 0.1$) en la tabla para ahorrar espacio, pero están listados en la tabla 11.3. Los resultados para w_i ($h = 0.025$) son exactos en aproximadamente 3×10^{-6}. Sin embargo, los resultados de Ect_{3i} son correctos para los lugares decimales enumerados. De hecho, si se usan suficientes dígitos, esta aproximación concordaría con la solución exacta con error máximo de 6.3×10^{-11} en los puntos de malla, una mejora impresionante. ■

La sección Conjunto de ejercicios 11.3 está disponible en línea. Encuentre la ruta de acceso en las páginas preliminares.

11.4 Métodos de diferencias finitas para problemas no lineales

Para el problema no lineal general de valor en la frontera

$$y'' = f(x, y, y'), \quad \text{para } a \leq x \leq b, \text{con } y(a) = \alpha \text{ y } y(b) = \beta,$$

el método de diferencias es similar al método aplicado para problemas lineales en la sección 11.3. Sin embargo, aquí el sistema de ecuaciones no será lineal, por lo que se requiere un proceso iterativo para resolverlo.

Para el desarrollo del procedimiento, suponemos siempre que f satisface las siguientes condiciones:

- f y las derivadas parciales f_y y $f_{y'}$ son todas continuas en

$$D = \{ (x, y, y') \mid a \leq x \leq b, \text{con } -\infty < y < \infty \text{ y } -\infty < y' < \infty \};$$

- $f_y(x, y, y') \geq \delta$ en D, para algún $\delta > 0$;
- Existen constantes k y L, con

$$k = \max_{(x,y,y') \in D} |f_y(x, y, y')| \quad \text{y} \quad L = \max_{(x,y,y') \in D} |f_{y'}(x, y, y')|.$$

Esto garantiza, mediante el teorema 11.1, que existe una solución única.

Como en el caso lineal, dividimos $[a, b]$ en $(N + 1)$ subintervalos iguales, cuyos extremos están en $x_i = a + ih$, para $i = 0, 1, \ldots, N + 1$. Suponer que la solución exacta tiene una cuarta derivada en la frontera nos permite reemplazar $y''(x_i)$ y $y'(x_i)$ en cada una de las ecuaciones

$$y''(x') = f(x', y(x\ ), y'(x'))$$

mediante la fórmula de diferencia centrada adecuada provista en las ecuaciones (11.16) y (11.17) en la página 518, respectivamente. Esto da, para cada $i = 1, 2, \ldots, N$,

$$\frac{y(x_{i+1}) - 2y(x_i) + y(x_{i-1})}{h^2} = f\left(x_i, y(x_i), \frac{y(x_{i+1}) - y(x_{i-1})}{2h} - \frac{h^2}{6} y'''(\eta_i)\right) + \frac{h^2}{12} y^{(4)}(\xi_i),$$

para algunas ξ_i y η_i en el intervalo (x_{i-1}, x_{i+1}).

Como en el caso lineal, el método de diferencias resulta de eliminar los términos de error y emplear las condiciones de frontera:

$$w_0 = \alpha, \quad w_{N+1} = \beta,$$

y

$$-\frac{w_{i+1}-2w_i+w_{i-1}}{h^2}+f\left(x_i, w_i, \frac{w_{i+1}-w_{i-1}}{2h}\right)=0,$$

para cada $i = 1, 2, \ldots, N$.

El sistema no lineal $N \times N$ obtenido a partir de este método,

$$\begin{aligned}
2w_1 - w_2 + h^2 f\left(x_1, w_1, \frac{w_2-\alpha}{2h}\right) - \alpha &= 0,\\
-w_1 + 2w_2 - w_3 + h^2 f\left(x_2, w_2, \frac{w_3-w_1}{2h}\right) &= 0,\\
&\vdots \\
-w_{N-2} + 2w_{N-1} - w_N + h^2 f\left(x_{N-1}, w_{N-1}, \frac{w_N-w_{N-2}}{2h}\right) &= 0,\\
-w_{N-1} + 2w_N + h^2 f\left(x_N, w_N, \frac{\beta-w_{N-1}}{2h}\right) - \beta &= 0,
\end{aligned} \tag{11.20}$$

tiene una solución única siempre y cuando $h < 2/L$, como se muestra en [Keller, H], p. 86. También, consulte el ejercicio 7.

Método de Newton para iteraciones

Usamos el método de Newton para los sistemas no lineales que se analizaron en la sección 10.2 para aproximar la solución de este sistema. Se genera una sucesión de iteraciones $\left\{\left(w_1^{(k)}, w_2^{(k)}, \ldots, w_N^{(k)}\right)^t\right\}$ que converge a la solución del sistema (11.20), siempre que la aproximación inicial $\left(w_1^{(0)}, w_2^{(0)}, \ldots, w_N^{(0)}\right)^t$ sea suficientemente cercana a la solución $(w_1, w_2, \ldots, w_N)^t$ y la matriz jacobina para que el sistema sea no singular. Para el sistema (11.20), la matriz jacobina $J(w_1, \ldots, w_N)$ es tridiagonal con ij-ésima entrada

$$J(w_1,\ldots,w_N)_{ij} = \begin{cases} -1 + \dfrac{h}{2} f_{y'}\left(x_i, w_i, \dfrac{w_{i+1}-w_{i-1}}{2h}\right), & \text{para } i = j-1 \text{ y } j = 2, \ldots, N,\\ 2 + h^2 f_y\left(x_i, w_i, \dfrac{w_{i+1}-w_{i-1}}{2h}\right), & \text{para } i = j \text{ y } j = 1, \ldots, N,\\ -1 - \dfrac{h}{2} f_{y'}\left(x_i, w_i, \dfrac{w_{i+1}-w_{i-1}}{2h}\right), & \text{para } i = j+1 \text{ y } j = 1, \ldots, N-1, \end{cases}$$

donde $w_0 = \alpha$ y $w_{N+1} = \beta$.

El método de Newton para los sistemas no lineales requiere que en cada iteración se resuelva el sistema lineal $N \times N$

$$\begin{aligned}
&J(w_1, \ldots, w_N)(v_1, \ldots, v_n)^t\\
&\quad = -\Bigg(2w_1 - w_2 - \alpha + h^2 f\left(x_1, w_1, \frac{w_2-\alpha}{2h}\right),\\
&\qquad - w_1 + 2w_2 - w_3 + h^2 f\left(x_2, w_2, \frac{w_3-w_1}{2h}\right), \ldots,
\end{aligned}$$

$$-w_{N-2}+2w_{N-1}-w_N+h^2 f\left(x_{N-1},w_{N-1},\frac{w_N-w_{N-2}}{2h}\right),$$

$$-w_{N-1}+2w_N+h^2 f\left(x_N,w_N,\frac{\beta-w_{N-1}}{2h}\right)-\beta\Bigg)^t$$

para $v_1, v_2, \ldots, v_N$ ya que

$$w_i^{(k)}=w_i^{(k-1)}+v_i,\quad \text{para cada } i=1,2,\ldots,N.$$

Puesto que J es tridiagonal, esto no es un problema tan difícil como podría parecer. En particular, se puede aplicar la factorización de Crout, el algoritmo 6.7 en la página 316. El proceso se detalla en el algoritmo 11.4.

ALGORITMO

11.4

Diferencia finita no lineal

Para aproximar la solución del problema de valor en la frontera no lineal

$$y''=f(x,y,y'),\quad \text{para } a\le x\le b, \text{con } y(a)=\alpha \text{ y } y(b)=\beta:$$

ENTRADA extremos a, b; condiciones de frontera α, β; entero $N\ge 2$; tolerancia TOL; número máximo de iteraciones M.

SALIDA aproximaciones w_i para $y(x_i)$ para cada $i=0,1,\ldots,N+1$ o un mensaje que inca que se ha excedido el número máximo de iteraciones.

Paso 1 Determine $h=(b-a)/(N+1)$;
$w_0=\alpha$;
$w_{N+1}=\beta$.

Paso 2 Para $i=1,\ldots,N$ determine $w_i=\alpha+i\left(\dfrac{\beta-\alpha}{b-a}\right)h$.

Paso 3 Determine $k=1$.

Paso 4 Mientras $k\le M$ haga los pasos 5–l6.

Paso 5 Determine $x=a+h$;
$t=(w_2-\alpha)/(2h)$;
$a_1=2+h^2 f_y(x,w_1,t)$;
$b_1=-1+(h/2)f_{y'}(x,w_1,t)$;
$d_1=-\big(2w_1-w_2-\alpha+h^2 f(x,w_1,t)\big)$.

Paso 6 Para $i=2,\ldots,N-1$
determine $x=a+ih$;
$t=(w_{i+1}-w_{i-1})/(2h)$;
$a_i=2+h^2 f_y(x,w_i,t)$;
$b_i=-1+(h/2)f_{y'}(x,w_i,t)$;
$c_i=-1-(h/2)f_{y'}(x,w_i,t)$;
$d_i=-\big(2w_i-w_{i+1}-w_{i-1}+h^2 f(x,w_i,t)\big)$.

Paso 7 Determine $x=b-h$;
$t=(\beta-w_{N-1})/(2h)$;
$a_N=2+h^2 f_y(x,w_N,t)$;
$c_N=-1-(h/2)f_{y'}(x,w_N,t)$;
$d_N=-\big(2w_N-w_{N-1}-\beta+h^2 f(x,w_N,t)\big)$.

Paso 8 Determine $l_1 = a_1$; (*Los pasos* 8-12 *resuelven un sistema lineal tridiagonal mediante el algoritmo* 6.7.)
$u_1 = b_1/a_1$;
$z_1 = d_1/l_1$.

Paso 9 Para $i = 2, \ldots, N-1$ determine $l_i = a_i - c_i u_{i-1}$;
$u_i = b_i/l_i$;
$z_i = (d_i - c_i z_{i-1})/l_i$.

Paso 10 Determine $l_N = a_N - c_N u_{N-1}$;
$z_N = (d_N - c_N z_{N-1})/l_N$.

Paso 11 Determine $v_N = z_N$;
$w_N = w_N + v_N$.

Paso 12 Para $i = N-1, \ldots, 1$ determine $v_i = z_i - u_i v_{i+1}$;
$w_i = w_i + v_i$.

Paso 13 Si $\|\mathbf{v}\| \leq TOL$ entonces haga los pasos 14 y 15.

Paso 14 Para $i = 0, \ldots, N+1$ determine $x = a + ih$;
SALIDA (x, w_i).

Paso 15 PARE. (*El procedimiento fue exitoso*.)

Paso 16 Determine $k = k + 1$.

Paso 17 SALIDA ('Número máximo de iteraciones excedido');
(*El procedimiento no fue exitoso*.)
PARE. ■

Se puede mostrar (consulte [IK], p. 433) que este método de diferencia finita no lineal es de orden $O(h^2)$.

Se requiere una buena aproximación inicial cuando no se puede verificar que se cumplen las condiciones (1), (2) y (3) dadas al principio de esta presentación, por lo que se debe especificar una cota superior para el número de iteraciones y, si se excede, se considera una aproximación inicial nueva o una reducción de la longitud. A menos que exista información contradictoria, es razonable comenzar el procedimiento al asumir que la solución es lineal. Por lo que, las aproximaciones iniciales $w_i^{(0)}$ para w_i, para cada $i = 1, 2, \ldots, N$, se obtiene en el paso 2 al pasar una línea recta a través de los extremos conocidos (a, α) y (b, β) y al evaluar en x_i.

Ejemplo 1 Aplique el algoritmo 11.4, con $h = 0.1$, para el problema no lineal de valor en la frontera

$$y'' = \frac{1}{8}(32 + 2x^3 - yy'), \quad \text{para } 1 \leq x \leq 3, \text{con } y(1) = 17 \text{ y } y(3) = \frac{43}{3},$$

y compare los resultados con los obtenidos en el ejemplo 1 de la sección 11.2.

Solución El procedimiento de paro o detención utilizado en el algoritmo 11.4 fue iterar hasta que los valores de iteraciones sucesivas difieran por menos de 10^{-8}. Se logró con cuatro iteraciones. Esto da los resultados en la tabla 11.5. Son menos exactos que los obtenidos usando el método de disparo no lineal, lo cual dio los resultados de la parte media de la tabla, exactos en el orden de 10^{-5}. ■

Tabla 11.5

x_i	w_i	$y(x_i)$	$\lvert w_i - y(x_i)\rvert$
1.0	17.000000	17.000000	
1.1	15.754503	15.755455	9.520×10^{-4}
1.2	14.771740	14.773333	1.594×10^{-3}
1.3	13.995677	13.997692	2.015×10^{-3}
1.4	13.386297	13.388571	2.275×10^{-3}
1.5	12.914252	12.916667	2.414×10^{-3}
1.6	12.557538	12.560000	2.462×10^{-3}
1.7	12.299326	12.301765	2.438×10^{-3}
1.8	12.126529	12.128889	2.360×10^{-3}
1.9	12.028814	12.031053	2.239×10^{-3}
2.0	11.997915	12.000000	2.085×10^{-3}
2.1	12.027142	12.029048	1.905×10^{-3}
2.2	12.111020	12.112727	1.707×10^{-3}
2.3	12.245025	12.246522	1.497×10^{-3}
2.4	12.425388	12.426667	1.278×10^{-3}
2.5	12.648944	12.650000	1.056×10^{-3}
2.6	12.913013	12.913846	8.335×10^{-4}
2.7	13.215312	13.215926	6.142×10^{-4}
2.8	13.553885	13.554286	4.006×10^{-4}
2.9	13.927046	13.927241	1.953×10^{-4}
3.0	14.333333	14.333333	

Uso de la extrapolación de Richardson

El procedimiento de extrapolación de Richardson también se puede usar para el método de diferencias finitas no lineal. La tabla 11.6 enumera los resultados cuando se aplica este método a nuestro ejemplo usando $h = 0.1, 0.05$, y 0.025,), con cuatro iteraciones en cada caso. Se omiten los valores de $w_i(h = 0.1)$ de la tabla para ahorrar espacio, pero se listan en la tabla 11.5. Los valores de $w_i(h = 0.25)$ son exactos dentro de aproximadamente 1.5×10^{-4}. Sin embargo, los valores de Ext_{3i} son exactos para los lugares enumerados, con un error máximo actual de 3.68×10^{-10}.

Tabla 11.6

x_i	$w_i(h = 0.05)$	$w_i(h = 0.025)$	Ext_{1i}	Ext_{2i}	Ext_{3i}
1.0	17.00000000	17.00000000	17.00000000	17.00000000	17.00000000
1.1	15.75521721	15.75539525	15.75545543	15.75545460	15.75545455
1.2	14.77293601	14.77323407	14.77333479	14.77333342	14.77333333
1.3	13.99718996	13.99756690	13.99769413	13.99769242	13.99769231
1.4	13.38800424	13.38842973	13.38857346	13.38857156	13.38857143
1.5	12.91606471	12.91651628	12.91666881	12.91666680	12.91666667
1.6	12.55938618	12.55984665	12.56000217	12.56000014	12.56000000
1.7	12.30115670	12.30161280	12.30176684	12.30176484	12.30176471
1.8	12.12830042	12.12874287	12.12899094	12.12888902	12.12888889
1.9	12.03049438	12.03091316	12.03105457	12.03105275	12.03105263
2.0	11.99948020	11.99987013	12.00000179	12.00000011	12.00000000
2.1	12.02857252	12.02892892	12.02902924	12.02904772	12.02904762
2.2	12.11230149	12.11262089	12.11272872	12.11272736	12.11272727
2.3	12.24614846	12.24642848	12.24652299	12.24652182	12.24652174
2.4	12.42634789	12.42658702	12.42666773	12.42666673	12.42666667
2.5	12.64973666	12.64993420	12.65000086	12.65000005	12.65000000
2.6	12.91362828	12.91379422	12.91384683	12.91384620	12.91384615
2.7	13.21577275	13.21588765	13.21592641	13.21592596	13.21592593
2.8	13.55418579	13.55426075	13.55428603	13.55428573	13.55428571
2.9	13.92719268	13.92722921	13.92724153	13.92724139	13.92724138
3.0	14.33333333	14.33333333	14.33333333	14.33333333	14.33333333

La sección Conjunto de ejercicios 11.4 está disponible en línea. Encuentre la ruta de acceso en las páginas preliminares.



11.5 El método de Rayleigh-Ritz

John William Strutt Lord Rayleigh (1842–1919), un físico matemático que se interesaba especialmente en la propagación de ondas, recibió el Premio Nobel de física en 1904.

Walter Ritz (1878–1909), un físico teórico en la Göttigen University, publicó un artículo sobre un problema de variación en 1909 [Ri]. Murió de tuberculosis a los 31 años.

El método del disparo para aproximar la solución de un problema de valor en la frontera reemplazó al problema de valor en la frontera con un par de problemas de valor inicial. El enfoque de diferencia finita reemplaza la operación continua de diferenciación con la operación discreta de diferencias finitas. El método Rayleigh-Ritz es una técnica de variación que aborda el problema desde un tercer enfoque. El problema de valor en la frontera primero se reformula como un problema de selección; del conjunto de todas las funciones suficientemente diferenciables que satisfacen las condiciones de frontera, se selecciona la función para minimizar cierta integral. A continuación, se reduce el tamaño del conjunto de funciones viables y se encuentra una aproximación a partir de este conjunto para minimizar la integral. Esto nos proporciona una aproximación para la solución del problema de valor en la frontera.

Para describir el método Rayleigh-Ritz consideramos la aproximación de una solución solución de un problema de valor en la frontera en dos puntos lineal desde el análisis de tensión de una viga. Este problema de valor en la frontera se describe mediante la ecuación diferencial

$$-\frac{d}{dx}\left(p(x)\frac{dy}{dx}\right)+q(x)y=f(x), \quad \text{para } 0\leq x\leq 1, \tag{11.21}$$

con las condiciones de frontera

$$y(0)=y(1)=0. \tag{11.22}$$

Esta ecuación diferencial describe la deflexión $y(x)$ de una viga de longitud 1 con sección transversal variable representada por $q(x)$. La deflexión se debe a las tensiones añadidas $p(x)$ y $f(x)$. En los ejercicios 9 y 12 se consideran condiciones de frontera más generales.

En el siguiente análisis suponemos que $p\in C^1[0,1]$ y $q, f\in C[0,1]$. Además, suponemos que existe una constante $\delta>0$ tal que

$$p(x)\geq\delta, \quad \text{y que } q(x)\geq 0, \quad \text{para cada } x \text{ en } [0,1].$$

Estas suposiciones son suficientes para garantizar que el problema de valor en la frontera provisto en las ecuaciones (11.21) y (11.22) tiene una solución única (consulte [BSW]).

Problemas de variaciones

Al igual que en el caso de los problemas de valor en la frontera que describen fenómenos físicos, la solución de la ecuación de la viga satisface la propiedad **variacional** de minimización integral. El principio variacional para la ecuación de la viga es fundamental para el desarrollo del método Rayleigh-Ritz y caracteriza la solución para dicha ecuación como la función que minimiza una integral sobre todas las funciones $C_0^2[0,1]$, el conjunto de esas funciones u en $C^2[0,1]$ con la propiedad $u(0)=u(1)=0$. El siguiente teorema establece la caracterización.

Teorema 11.4 Si $p\in C^1[0,1]$, $q, f\in C[0,1]$, y

$$p(x)\geq\delta>0, \quad q(x)\geq 0, \quad \text{para } 0\leq x\leq 1.$$

La función $y\in C_0^2[0,1]$ es la única solución para la ecuación diferencial

$$-\frac{d}{dx}\left(p(x)\frac{dy}{dx}\right)+q(x)y=f(x), \quad \text{para } 0\leq x\leq 1, \tag{11.23}$$

si y sólo si y es la única función en $C_0^2[0,1]$ que minimiza la integral

$$I[u]=\int_0^1\{p(x)[u'(x)]^2+q(x)[u(x)]^2-2f(x)u(x)\}\,dx. \tag{11.24}$$

■

Los detalles de la demostración de este teorema se pueden encontrar en [Shul], p. 88-89. Su procedimiento consta de tres pasos. Primero se muestra que una función, y, es una solución de la solución de la ecuación (11.23) si y sólo si satisface la ecuación

- $$\int_0^1 f(x)u(x)dx = \int_0^1 p(x)y'(x)u'(x) + q(x)y(x)u(x)dx, \tag{11.25}$$
  para toda $u \in C_0^2[0.1]$.
- El segundo paso muestra que $y \in C_0^2[0, 1]$ es una solución para la ecuación (11.24) si y sólo si la ecuación (11.25) se mantiene para toda $u \in C_0^2[0, 1]$.
- El paso final muestra que (11.5) tiene una solución única. Esta última también será una solución para (11.24) y para (11.23), de tal forma que las soluciones para las ecuaciones (11.23) y (11.24) son idénticas.

El método Rayleigh-Ritz aproxima la solución y al minimizar la integral no sobre todas las funciones en $C_0^2[0, 1]$, sino sobre un conjunto pequeño de funciones que consisten en combinaciones lineales de ciertas funciones base $\phi_1, \phi_2, \ldots, \phi_n$. Las funciones base son linealmente independientes y satisfacen

$$\phi_i(0) = \phi_i(1) = 0, \quad \text{para cada } i = 1, 2, \ldots, n.$$

A continuación se obtiene una aproximación $\phi(x) = \sum_{i=1}^n c_i\phi_i(x)$ para la solución $y(x)$ de la ecuación (11.23) al encontrar las constantes $c_1, c_2, \ldots, c_n$ para minimizar la integral $I\left[\sum_{i=1}^n c_i\phi_i\right]$.

A partir de la ecuación (11.24)

$$I[\phi] = I\left[\sum_{i=1}^n c_i\phi_i\right] \tag{11.26}$$

$$= \int_0^1 \left\{ p(x)\left[\sum_{i=1}^n c_i\phi_i'(x)\right]^2 + q(x)\left[\sum_{i=1}^n c_i\phi_i(x)\right]^2 - 2f(x)\sum_{i=1}^n c_i\phi_i(x) \right\} dx,$$

y, para que se presente un mínimo, es necesario, al considerar I como una función de $c_1, c_2, \ldots, c_n$, tener

$$\frac{\partial I}{\partial c_j} = 0, \quad \text{para cada } j = 1, 2, \ldots, n. \tag{11.27}$$

Al derivar (11.26) obtenemos

$$\frac{\partial I}{\partial c_j} = \int_0^1 \left\{ 2p(x)\sum_{i=1}^n c_i\phi_i'(x)\phi_j'(x) + 2q(x)\sum_{i=1}^n c_i\phi_i(x)\phi_j(x) - 2f(x)\phi_j(x) \right\} dx,$$

y al sustituir en la ecuación (11.27) se obtiene

$$0 = \sum_{i=1}^n \left[\int_0^1 \{p(x)\phi_i'(x)\phi_j'(x) + q(x)\phi_i(x)\phi_j(x)\}\, dx\right] c_i - \int_0^1 f(x)\phi_j(x)\, dx, \tag{11.28}$$

para cada $j = 1, 2, \ldots, n$.

Las **ecuaciones normales** descritas en la ecuación (11.28) generan un sistema lineal $n \times n$ $A\mathbf{c} = \mathbf{b}$ en las variables $c_1, c_2, \ldots, c_n$, donde la matriz simétrica A está dada

$$a_{ij} = \int_0^1 [p(x)\phi_i'(x)\phi_j'(x) + q(x)\phi_i(x)\phi_j(x)]\, dx,$$

y **b** se define como

$$b_i = \int_0^1 f(x)\phi_i(x)\, dx.$$

Base lineal por tramos

La opción más simple de las funciones base implica polinomios lineales por tramos. El primer paso es formar una partición de [0, 1] al seleccionar los puntos $x_0, x_1, \ldots, x_{n+1}$ con

$$0 = x_0 < x_1 < \cdots < x_n < x_{n+1} = 1.$$

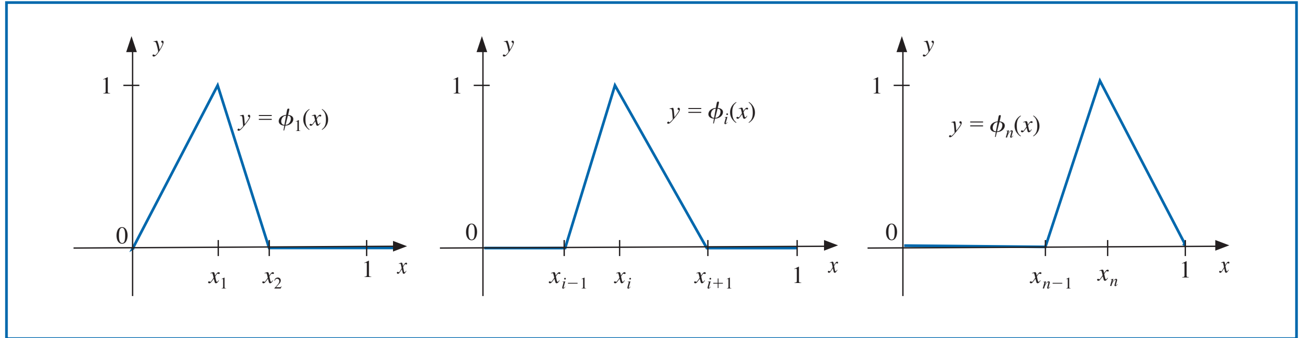
Al utilizar $h_i = x_{i+1} - x_i$, para cada $i = 0, 1, \ldots, n$, definimos las funciones base $\phi_1(x)$, $\phi_2(x), \ldots, \phi_n(x)$ mediante

$$\phi_i(x) = \begin{cases} 0, & \text{si} \quad 0 \le x \le x_{i-1}, \\ \dfrac{1}{h_{i-1}}(x - x_{i-1}), & \text{si} \quad x_{i-1} < x \le x_i, \\ \dfrac{1}{h_i}(x_{i+1} - x), & \text{si} \quad x_i < x \le x_{i+1}, \\ 0, & \text{si} \quad x_{i+1} < x \le 1, \end{cases} \tag{11.29}$$

para cada $i = 0, 1, \ldots, n$. (Consulte la figura 11.4.)

En el ejercicio 13 se muestra que las funciones base son linealmente independientes.

Figura 11.4



Las funciones ϕ_i son lineales por tramos, por lo que las derivadas ϕ_i', aunque no sean continuas, son constantes en (x_j, x_{j+1}), para cada $j = 0, 1, \ldots, n$, y

$$\phi_i'(x) = \begin{cases} 0, & \text{si} \quad 0 < x < x_{i-1}, \\ \dfrac{1}{h_{i-1}}, & \text{si} \quad x_{i-1} < x < x_i, \\ -\dfrac{1}{h_i}, & \text{si} \quad x_i < x < x_{i+1}, \\ 0, & \text{si} \quad x_{i+1} < x < 1, \end{cases} \tag{11.30}$$

para cada $i = 1, 2, \ldots, n$;

Puesto que ϕ_i y ϕ_i' son diferentes a cero sólo en (x_{i-1}, x_{i+1}),

$$\phi_i(x)\phi_j(x) \equiv 0 \qquad \text{y} \qquad \phi_i'(x)\phi_j'(x) \equiv 0,$$

excepto cuando j es $i-1$, i, o $i+1$. Como consecuencia, el sistema lineal dado por la ecuación (11.28) se reduce a un sistema lineal tridiagonal $n \times n$. Las entradas diferentes a cero en A son

$$\begin{aligned}
a_{ii} &= \int_0^1 \left\{ p(x)[\phi_i'(x)]^2 + q(x)[\phi_i(x)]^2 \right\} dx \\
&= \left(\frac{1}{h_{i-1}}\right)^2 \int_{x_{i-1}}^{x_i} p(x)\, dx + \left(\frac{-1}{h_i}\right)^2 \int_{x_i}^{x_{i+1}} p(x)\, dx \\
&\quad + \left(\frac{1}{h_{i-1}}\right)^2 \int_{x_{i-1}}^{x_i} (x - x_{i-1})^2 q(x)\, dx + \left(\frac{1}{h_i}\right)^2 \int_{x_i}^{x_{i+1}} (x_{i+1} - x)^2 q(x)\, dx,
\end{aligned}$$

para cada $i = 1, 2, \ldots, n$;

$$\begin{aligned}
a_{i,i+1} &= \int_0^1 \{p(x)\phi_i'(x)\phi_{i+1}'(x) + q(x)\phi_i(x)\phi_{i+1}(x)\}\, dx \\
&= -\left(\frac{1}{h_i}\right)^2 \int_{x_i}^{x_{i+1}} p(x)\, dx + \left(\frac{1}{h_i}\right)^2 \int_{x_i}^{x_{i+1}} (x_{i+1} - x)(x - x_i) q(x)\, dx,
\end{aligned}$$

para cada $i = 1, 2, \ldots, n-1$; y

$$\begin{aligned}
a_{i,i-1} &= \int_0^1 \{p(x)\phi_i'(x)\phi_{i-1}'(x) + q(x)\phi_i(x)\phi_{i-1}(x)\}\, dx \\
&= -\left(\frac{1}{h_{i-1}}\right)^2 \int_{x_{i-1}}^{x_i} p(x)\, dx + \left(\frac{1}{h_{i-1}}\right)^2 \int_{x_{i-1}}^{x_i} (x_i - x)(x - x_{i-1}) q(x)\, dx,
\end{aligned}$$

para cada $i = 2, \ldots, n$. Las entradas de $\mathbf{b}$ son

$$b_i = \int_0^1 f(x)\phi_i(x)\, dx = \frac{1}{h_{i-1}} \int_{x_{i-1}}^{x_i} (x - x_{i-1}) f(x)\, dx + \frac{1}{h_i} \int_{x_i}^{x_{i+1}} (x_{i+1} - x) f(x)\, dx,$$

para cada $i = 1, 2, \ldots, n$.

Existen seis tipos de integrales que deben evaluarse:

$$Q_{1,i} = \left(\frac{1}{h_i}\right)^2 \int_{x_i}^{x_{i+1}} (x_{i+1} - x)(x - x_i) q(x)\, dx, \quad \text{para cada} \quad i = 1, 2, \ldots, n-1,$$

$$Q_{2,i} = \left(\frac{1}{h_{i-1}}\right)^2 \int_{x_{i-1}}^{x_i} (x - x_{i-1})^2 q(x)\, dx, \quad \text{para cada} \quad i = 1, 2, \ldots, n,$$

$$Q_{3,i} = \left(\frac{1}{h_i}\right)^2 \int_{x_i}^{x_{i+1}} (x_{i+1} - x)^2 q(x)\, dx, \quad \text{para cada} \quad i = 1, 2, \ldots, n,$$

$$Q_{4,i} = \left(\frac{1}{h_{i-1}}\right)^2 \int_{x_{i-1}}^{x_i} p(x)\, dx, \quad \text{para cada} \quad i = 1, 2, \ldots, n+1,$$

$$Q_{5,i} = \frac{1}{h_{i-1}} \int_{x_{i-1}}^{x_i} (x - x_{i-1}) f(x)\, dx, \quad \text{para cada} \quad i = 1, 2, \ldots, n,$$

y

$$Q_{6,i} = \frac{1}{h_i} \int_{x_i}^{x_{i+1}} (x_{i+1} - x) f(x)\, dx, \quad \text{para cada} \quad i = 1, 2, \ldots, n.$$

La matriz A y el vector $\mathbf{b}$ en el sistema lineal $A\mathbf{c} = \mathbf{b}$ tiene las entradas

$$\begin{aligned} a_{i,i} &= Q_{4,i} + Q_{4,i+1} + Q_{2,i} + Q_{3,i}, \quad \text{para cada} \quad i = 1, 2, \ldots, n, \\ a_{i,i+1} &= -Q_{4,i+1} + Q_{1,i}, \quad \text{para cada} \quad i = 1, 2, \ldots, n-1, \\ a_{i,i-1} &= -Q_{4,i} + Q_{1,i-1}, \quad \text{para cada} \quad i = 2, 3, \ldots, n, \end{aligned}$$

y

$$b_i = Q_{5,i} + Q_{6,i}, \quad \text{para cada} \quad i = 1, 2, \ldots, n.$$

Las entradas en $\mathbf{c}$ son los coeficientes desconocidos $c_1, c_2, \ldots, c_n$, a partir de los cuales se construye la aproximación de Rayleigh-Ritz ϕ, dada por $\phi(x) = \sum_{i=1}^{n} c_i \phi_i(x)$,

El uso de este método requiere que se evalúen $6n$ integrales, ya sea directamente o mediante una fórmula de cuadratura, como la regla compuesta de Simpson.

Un enfoque alternativo para la evaluación integral es aproximar cada una de las funciones *p, q* y *f* con su polinomio de interpolación lineal por tramos y, a continuación, integramos la aproximación.

Considere, por ejemplo, la integral $Q_{1,i}$. La interpolación lineal por tramos de q es

$$P_q(x) = \sum_{i=0}^{n+1} q(x_i)\phi_i(x),$$

donde $\phi_1, \ldots, \phi_n$ se definen en la ecuación(11.30) y

$$\phi_0(x) = \begin{cases} \dfrac{x_1 - x}{x_1}, & \text{si} \quad \le x \le x_1 \\ 0, & \text{en otro caso} \end{cases} \qquad \text{y} \qquad \phi_{n+1}(x) = \begin{cases} \dfrac{x - x_n}{1 - x_n}, & \text{si} \quad x_n \le x \le 1 \\ 0, & \text{en otro caso.} \end{cases}$$

El intervalo de integración es $[x_i, x_{i+1}]$, de tal forma que el polinomio por tramos $P_q(x)$ se reduce a

$$P_q(x) = q(x_i)\phi_i(x) + q(x_{i+1})\phi_{i+1}(x).$$

Éste es el polinomio interpolante de primer grado en la sección 3.1. Mediante el teorema 3.3 en la página 83,

$$|q(x) - P_q(x)| = O(h_i^2), \quad \text{para } x_i \le x \le x_{i+1},$$

si $q \in C^2[x_i, x_{i+1}]$. Para $i = 1, 2, \ldots, n-1$, la aproximación para $Q_{1,i}$ se obtiene al integrar la aproximación para el integrando

$$\begin{aligned} Q_{1,i} &= \left(\frac{1}{h_i}\right)^2 \int_{x_i}^{x_{i+1}} (x_{i+1} - x)(x - x_i) q(x)\, dx \\ &\approx \left(\frac{1}{h_i}\right)^2 \int_{x_i}^{x_{i+1}} (x_{i+1} - x)(x - x_i) \left[\frac{q(x_i)(x_{i+1} - x)}{h_i} + \frac{q(x_{i+1})(x - x_i)}{h_i} \right] dx \\ &= \frac{h_i}{12}[q(x_i) + q(x_{i+1})]. \end{aligned}$$

Además, si $q \in C^2[x_i, x_{i+1}]$, entonces

$$\left| Q_{1,i} - \frac{h_i}{12}[q(x_i) + q(x_{i+1})] \right| = O(h_i^3).$$

Las aproximaciones para las otras integrales se derivan de forma similar y están dadas por

$$Q_{2,i} \approx \frac{h_{i-1}}{12}[3q(x_i) + q(x_{i-1})], \qquad Q_{3,i} \approx \frac{h_i}{12}[3q(x_i) + q(x_{i+1})],$$

$$Q_{4,i} \approx \frac{1}{2h_{i-1}}[p(x_i) + p(x_{i-1})], \qquad Q_{5,i} \approx \frac{h_{i-1}}{6}[2f(x_i) + f(x_{i-1})],$$

y

$$Q_{6,i} \approx \frac{h_i}{6}[2f(x_i) + f(x_{i+1})].$$

El algoritmo 11.5 establece el sistema lineal tridiagonal e incluye el algoritmo de factorización de Crout 6.7 para resolver el sistema. Las integrales $Q_{1,i}, \ldots, Q_{6,i}$ se pueden calcular mediante alguno de los métodos antes mencionados.

ALGORITMO **11.5**

Rayleigh-Ritz lineal por tramos

Para aproximar la solución al problema de valor en la frontera

$$-\frac{d}{dx}\left(p(x)\frac{dy}{dx}\right) + q(x)y = f(x), \quad \text{para } 0 \le x \le 1, \text{con } y(0) = 0 \text{ y } y(1) = 0$$

con la función lineal por tramos

$$\phi(x) = \sum_{i=1}^{n} c_i \phi_i(x):$$

ENTRADA entero $n \ge 1$; puntos $x_0 = 0 < x_1 < \cdots < x_n < x_{n+1} = 1$.

SALIDA coeficientes $c_1, \ldots, c_n$.

Paso 1 Para $i = 0, \ldots, n$ determine $h_i = x_{i+1} - x_i$.

Paso 2 Para $i = 1, \ldots, n$ definir la base lineal por tramos ϕ_i por

$$\phi_i(x) = \begin{cases} 0, & 0 \le x \le x_{i-1}, \\ \dfrac{x - x_{i-1}}{h_{i-1}}, & x_{i-1} < x \le x_i, \\ \dfrac{x_{i+1} - x}{h_i}, & x_i < x \le x_{i+1}, \\ 0, & x_{i+1} < x \le 1. \end{cases}$$

Paso 3 Para cada $i = 1, 2, \ldots, n-1$ calcule $Q_{1,i}, Q_{2,i}, Q_{3,i}, Q_{4,i}, Q_{5,i}, Q_{6,i}$;
Calcule $Q_{2,n}, Q_{3,n}, Q_{4,n}, Q_{4,n+1}, Q_{5,n}, Q_{6,n}$.

Paso 4 Para cada $i = 1, 2, \ldots, n-1$, determine $\alpha_i = Q_{4,i} + Q_{4,i+1} + Q_{2,i} + Q_{3,i}$;
$\beta_i = Q_{1,i} - Q_{4,i+1}$;
$b_i = Q_{5,i} + Q_{6,i}$.

Paso 5 Determine $\alpha_n = Q_{4,n} + Q_{4,n+1} + Q_{2,n} + Q_{3,n}$;
$b_n = Q_{5,n} + Q_{6,n}$.

Paso 6 Determine $a_1 = \alpha_1$; (*Los pasos* 1-6 *resuelven el sistema lineal tridiagonal simétrico usando el algoritmo* 6.7.)
$\zeta_1 = \beta_1/\alpha_1$;
$z_1 = b_1/a_1$.

Paso 7 Para $i = 2, \ldots, n-1$ determine $a_i = \alpha_i - \beta_{i-1}\zeta_{i-1}$;
$\zeta_i = \beta_i / a_i$;
$z_i = (b_i - \beta_{i-1} z_{i-1})/a_i$.

Paso 8 Determine $a_n = \alpha_n - \beta_{n-1}\zeta_{n-1}$;
$z_n = (b_n - \beta_{n-1} z_{n-1})/a_n$.

Paso 9 Determine $c_n = z_n$;
SALIDA (c_n).

Paso 10 Para $i = n-1, \ldots, 1$ determine $c_i = z_i - \zeta_i c_{i+1}$;
SALIDA (c_i).

Paso 11 PARE. (*El procedimiento está completo*.) ■

La siguiente ilustración usa el algoritmo 11.5. Debido a la naturaleza fundamental de este ejemplo, las integrales en los pasos 3, 4 y 5 se encontraron directamente.

Ilustración Considere el problema de valor en la frontera

$$-y'' + \pi^2 y = 2\pi^2 \operatorname{sen}(\pi x), \quad \text{para } 0 \le x \le 1, \text{ con } y(0) = y(1) = 0.$$

Sea $h_i = h = 0.1$, de tal forma que $x_i = 0.1i$, para cada $i = 0, 1, \ldots, 9$. Las integrales son

$$Q_{1,i} = 100 \int_{0.1i}^{0.1i+0.1} (0.1i + 0.1 - x)(x - 0.1i)\pi^2 \, dx = \frac{\pi^2}{60},$$

$$Q_{2,i} = 100 \int_{0.1i-0.1}^{0.1i} (x - 0.1i + 0.1)^2 \pi^2 \, dx = \frac{\pi^2}{30},$$

$$Q_{3,i} = 100 \int_{0.1i}^{0.1i+0.1} (0.1i + 0.1 - x)^2 \pi^2 \, dx = \frac{\pi^2}{30},$$

$$Q_{4,i} = 100 \int_{0.1i-0.1}^{0.1i} dx = 10,$$

$$\begin{aligned} Q_{5,i} &= 10 \int_{0.1i-0.1}^{0.1i} (x - 0.1i + 0.1) 2\pi^2 \operatorname{sen} \pi x \, dx \\ &= -2\pi \cos 0.1\pi i + 20[\operatorname{sen}(0.1\pi i) - \operatorname{sen}((0.1i - 0.1)\pi)], \end{aligned}$$

y

$$\begin{aligned} Q_{6,i} &= 10 \int_{0.1i}^{0.1i+0.1} (0.1i + 0.1 - x) 2\pi^2 \operatorname{sen} \pi x \, dx \\ &= 2\pi \cos 0.1\pi i - 20[\operatorname{sen}((0.1i + 0.1)\pi) - \operatorname{sen}(0.1\pi i)]. \end{aligned}$$

El sistema lineal $A\mathbf{c} = \mathbf{b}$ tiene

$$a_{i,i} = 20 + \frac{\pi^2}{15}, \qquad \text{para cada } i = 1, 2, \ldots, 9,$$

$$a_{i,i+1} = -10 + \frac{\pi^2}{60}, \qquad \text{para cada } i = 1, 2, \ldots, 8,$$

$$a_{i,i-1} = -10 + \frac{\pi^2}{60}, \qquad \text{para cada } i = 2, 3, \ldots, 9,$$

y

$$b_i = 40\,\text{sen}(0.1\pi i)[1 - \cos 0.1\pi], \qquad \text{para cada } i = 1, 2, \ldots, 9.$$

La solución para el sistema lineal tridiagonal es

$$c_9 = 0.3102866742,\ c_8 = 0.5902003271,\ c_7 = 0.8123410598,$$
$$c_6 = 0.9549641893,\ c_5 = 1.004108771,\ \ c_4 = 0.9549641893,$$
$$c_3 = 0.8123410598,\ c_2 = 0.5902003271,\ c_1 = 0.3102866742.$$

La aproximación lineal por tramos es

$$\phi(x) = \sum_{i=1}^{9} c_i \phi_i(x),$$

y la solución real para el problema de valor en la frontera es $y(x) = \text{sen}\,\pi x$. La tabla 11.7 incluye el error en la aproximación en x_i, para cada $i = 1, \ldots, 9$. ■

Tabla 11.7

i	x_i	$\phi(x_i)$	$y(x_i)$	$\lvert\phi(x_i) - y(x_i)\rvert$
1	0.1	0.3102866742	0.3090169943	0.00127
2	0.2	0.5902003271	0.5877852522	0.00241
3	0.3	0.8123410598	0.8090169943	0.00332
4	0.4	0.9549641896	0.9510565162	0.00390
5	0.5	1.0041087710	1.0000000000	0.00411
6	0.6	0.9549641893	0.9510565162	0.00390
7	0.7	0.8123410598	0.8090169943	0.00332
8	0.8	0.5902003271	0.5877852522	0.00241
9	0.9	0.3102866742	0.3090169943	0.00127

Se puede mostrar que la matriz tridiagonal A provista por las funciones base lineales por tramos es definida positiva (consulte el ejercicio 15), por lo que, por el teorema 6.26 en la página 311, el sistema lineal es estable respecto al error de redondeo. De acuerdo con la hipótesis presentada al inicio de esta sección, tenemos

$$|\phi(x) - y(x)| = O(h^2), \quad \text{para cada x en } [0, 1].$$

Una prueba de este resultado se puede encontrar en [Schul], p. 103-104.

Base B spline

El uso de las funciones base lineales por tramos resulta en una solución aproximada para las ecuaciones (11.22) y (11.23) que es continua, pero no diferenciable en [0, 1]. Se requiere un conjunto más sofisticado de funciones base $C_0^2[0, 1]$ para construir una aproximación. Estas funciones base son similares a los splines cúbicos interpolantes analizados en la sección 3.5.

Recuerde que el spline cúbico *interpolante* S en los cinco nodos x_0, x_1, x_2, x_3 y x_4 para una función f está definido por:

a) $S(x)$ es un polinomio cúbico, que se denota $S_j(x)$, en el subintervalo $[x_j, x_{j+1}]$ para cada $j = 0, 1, 2, 3, 4$;

b) $S_j(x_j) = f(x_j)$ y $S_j(x_{j+1}) = f(x_{j+1})$ para cada $j = 0, 1, 2$;

c) $S_{j+1}(x_{j+1}) = S_j(x_{j+1})$ para cada $j = 0, 1, 2$; (*implicado por* **(b)**.)

d) $S'_{j+1}(x_{j+1}) = S'_j(x_{j+1})$ para cada $j = 0, 1, 2$;

e) $S''_{j+1}(x_{j+1}) = S''_j(x_{j+1})$ para cada $j = 0, 1, 2$;

f) Se debe satisfacer uno de los siguientes conjuntos de condiciones de frontera:

i) $S''(x_0) = S''(x_n) = 0$ *(frontera natural (o libre))*

ii) $S'(x_0) = f'(x_0)$ y $S'(x_n) = f'(x_n)$ *(frontera fijada)*.

Puesto que la singularidad de la solución requiere que el número de constantes en **a)**, 16, sea igual al número de condiciones en **b)** a través de **f)**, sólo se puede especificar una de las condiciones de frontera en **f)** para los splines cúbicos interpolantes.

En 1946, I. J. Schoenberg [Scho] presentó los B-(para "base"), pero por más de una década fueron difíciles de calcular. En 1972 Carl de Boor (1937–) [Deb1] describió las fórmulas para la evaluación que mejoraron su estabilidad y utilidad.

Las funciones de spline cúbico que usaremos para nuestras funciones base reciben el nombre de **B-splines** o *splines con forma de campana*. Estos difieren de los splines interpolantes en que se satisfacen ambos conjuntos de condiciones de frontera en **(f)**. Esto requiere que las dos condiciones en **b**) a **e**) se flexibilicen. Puesto que el spline debe tener dos derivadas continuas en $[x_0, x_4]$, borramos dos de las condiciones de interpolación a partir de la descripción de los splines interpolantes. En especial, modificamos la condición **b)** por

b. $S(x_j) = f(x_j)$ para $j = 0, 2, 4$.

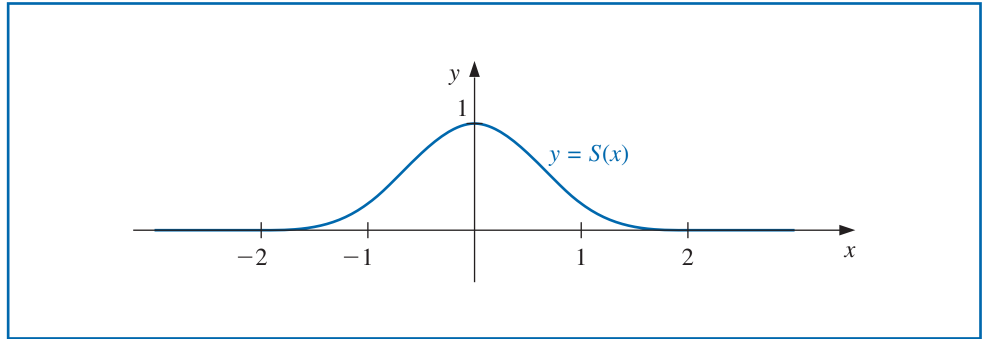
Por ejemplo, el B-spline básico S que se define a continuación y que se muestra en la figura 11.5 usa los nodos igualmente espaciados $x_0 = -2$, $x_1 = -1$, $x_2 = 0$, $x_3 = 1$ y $x_4 = 2$. Satisface las condiciones interpolantes

b. $S(x_0) = 0, \quad S(x_2) = 1, \quad S(x_4) = 0$

así como ambos conjuntos de condiciones

i) $S''(x_0) = S''(x_4) = 0$ y **ii)** $S'(x_0) = S'(x_4) = 0$.

Figura 11.5



Por consiguiente, $S \in C_0^2(-\infty, \infty)$, y está dado específicamente como

$$S(x) = \begin{cases} 0, & \text{si} \quad x \leq -2, \\ \frac{1}{4}(2+x)^3, & \text{si} \quad -2 \leq x \leq -1, \\ \frac{1}{4}\left[(2+x)^3 - 4(1+x)^3\right], & \text{si} \quad -1 < x \leq 0, \\ \frac{1}{4}\left[(2-x)^3 - 4(1-x)^3\right], & \text{si} \quad 0 < x \leq 1, \\ \frac{1}{4}(2-x)^3, & \text{si} \quad 1 < x \leq 2, \\ 0, & \text{si} \quad 2 < x. \end{cases} \tag{11.31}$$

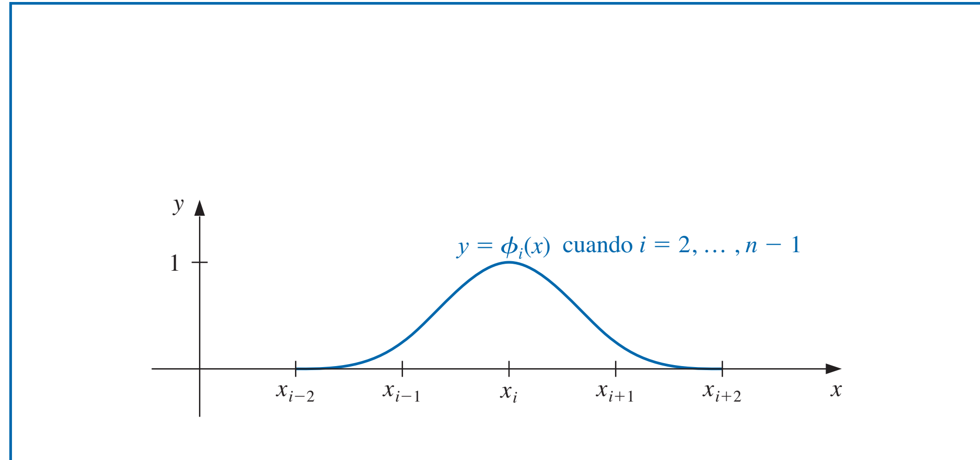
Ahora usaremos el B-spline básico para construir las funciones base ϕ_i en $C_0^2[0, 1]$. Primero dividimos $[0, 1]$ al seleccionar un entero positivo n y definir $h = 1/(n+1)$. Esto

produce nodos igualmente espaciados $x_i = ih$, para cada $i = 0, 1, \ldots, n+1$. Ahora definimos las funciones base $\{\phi_i\}_{i=0}^{n+1}$ como

$$\phi_i(x) = \begin{cases} S\left(\dfrac{x}{h}\right) - 4S\left(\dfrac{x+h}{h}\right), & \text{si} \quad i = 0, \\ S\left(\dfrac{x-h}{h}\right) - S\left(\dfrac{x+h}{h}\right), & \text{si} \quad i = 1, \\ S\left(\dfrac{x-ih}{h}\right), & \text{si} \quad 2 \le i \le n-1, \\ S\left(\dfrac{x-nh}{h}\right) - S\left(\dfrac{x-(n+2)h}{h}\right), & \text{si} \quad i = n, \\ S\left(\dfrac{x-(n+1)h}{h}\right) - 4S\left(\dfrac{x-(n+2)h}{h}\right), & \text{si} \quad i = n+1. \end{cases}$$

No es difícil mostrar que $\{\phi_i\}_{i=0}^{n+1}$ es un conjunto linealmente independiente de splines cúbicos que satisfacen $\phi_i(0) = \phi_i(1) = 0$, para cada $i = 0, 1, \ldots, n, n+1$ (consulte el ejercicio 14). Las gráficas de ϕ_i, para $2 \le i \le n-1$, se muestran en la figura 11.6 y las gráficas de ϕ_0, ϕ_1, ϕ_n, y ϕ_{n+1} están en la figura 11.7.

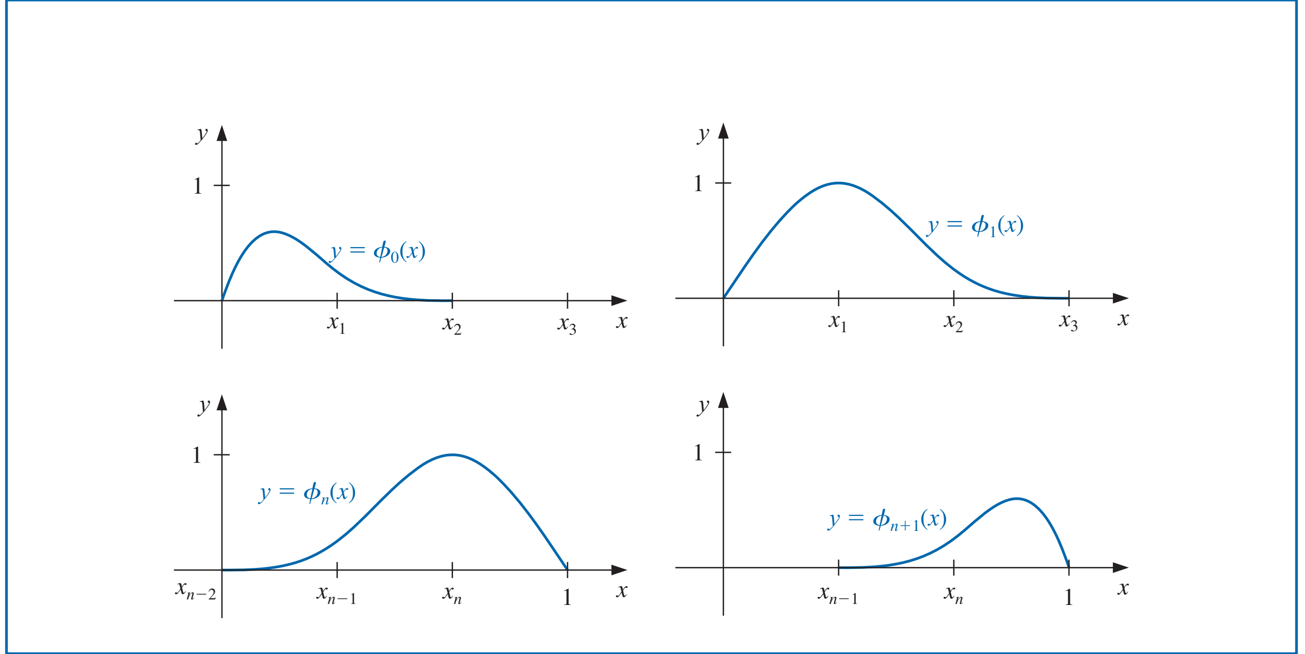
Figura 11.6



Puesto que $\phi_i(x)$ y $\phi_i'(x)$ son diferentes de cero sólo para $x \in [x_{i-2}, x_{i+2}]$, la matriz en la aproximación Rayleigh Ritz es una matriz de banda con ancho de banda máximo de siete:

$$A = \begin{bmatrix} a_{00} & a_{01} & a_{02} & a_{03} & 0 & \cdots & \cdots & 0 \\ a_{10} & a_{11} & a_{12} & a_{13} & a_{14} & \ddots & & \vdots \\ a_{20} & a_{21} & a_{22} & a_{23} & a_{24} & a_{25} & \ddots & \vdots \\ a_{30} & a_{31} & a_{32} & a_{33} & a_{34} & a_{35} & a_{36} & 0 \\ 0 & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & a_{n-2,n+1} \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & a_{n-1,n+1} \\ \vdots & & \ddots & \ddots & \ddots & \ddots & \ddots & a_{n,n+1} \\ 0 & \cdots & \cdots & 0 & a_{n+1,n-2} & a_{n+1,n-1} & a_{n+1,n} & a_{n+1,n+1} \end{bmatrix}$$

Figura 11.7



donde

$$a_{ij} = \int_0^1 \{p(x)\phi_i'(x)\phi_j'(x) + q(x)\phi_i(x)\phi_j(x)\}\, dx,$$

para cada $i, j = 0, 1, \ldots, n+1$. El vector $\mathbf{b}$ tiene las entradas

$$b_i = \int_0^1 f(x)\phi_i(x)dx.$$

La matriz A es definida positiva (consulte el ejercicio 16), por lo que el sistema lineal $A\mathbf{c} = \mathbf{b}$ se puede resolver mediante el algoritmo de Cholesky 6.6 o la eliminación gaussiana. El algoritmo 11.6 describe detalladamente la construcción de la aproximación del spline cúbico $\phi(x)$ mediante el método de Rayleigh-Ritz para los problemas de valor en la frontera (11.21) y (11.22) enunciados al principio de esta sección.

ALGORITMO **11.6**

Método Rayleigh-Ritz de spline cúbico

Para aproximar la solución del problema de valor en la frontera

$$-\frac{d}{dx}\left(p(x)\frac{dy}{dx}\right) + q(x)y = f(x), \quad \text{para } 0 \le x \le 1, \text{ con } y(0) = 0 \text{ y } y(1) = 0$$

con la suma de splines cúbicos

$$\phi(x) = \sum_{i=0}^{n+1} c_i\phi_i(x) :$$

ENTRADA entero $n \ge 1$.

SALIDA coeficientes $c_0, \ldots, c_{n+1}$.

Paso 1 Determine $h = 1/(n+1)$.

Paso 2 Para $i = 0, \ldots, n+1$ establezca $x_i = ih$.
Determine $x_{-2} = x_{-1} = 0$; $x_{n+2} = x_{n+3} = 1$.

Paso 3 Defina la función S mediante

$$S(x) = \begin{cases} 0, & x \leq -2, \\ \frac{1}{4}(2+x)^3, & -2 < x \leq -1, \\ \frac{1}{4}\left[(2+x)^3 - 4(1+x)^3\right], & -1 < x \leq 0, \\ \frac{1}{4}\left[(2-x)^3 - 4(1-x)^3\right], & 0 < x \leq 1, \\ \frac{1}{4}(2-x)^3, & 1 < x \leq 2, \\ 0, & 2 < x \end{cases}$$

Paso 4 Defina la base del spline cúbico $\{\phi_i\}_{i=0}^{n+1}$ mediante

$$\phi_0(x) = S\left(\frac{x}{h}\right) - 4S\left(\frac{x+h}{h}\right),$$

$$\phi_1(x) = S\left(\frac{x-x_1}{h}\right) - S\left(\frac{x+h}{h}\right),$$

$$\phi_i(x) = S\left(\frac{x-x_i}{h}\right), \text{ para } i = 2, \ldots, n-1,$$

$$\phi_n(x) = S\left(\frac{x-x_n}{h}\right) - S\left(\frac{x-(n+2)h}{h}\right),$$

$$\phi_{n+1}(x) = S\left(\frac{x-x_{n+1}}{h}\right) - 4S\left(\frac{x-(n+2)h}{h}\right).$$

Paso 5 Para $i = 0, \ldots, n+1$ haga los pasos 6–9.
(*Nota: Las integrales en los pasos* 6 *y* 9 *se pueden evaluar usando un procedimiento de integración numérica.*)

Paso 6 Para $j = i, i+1, \ldots, \text{mín}\{i+3, n+1\}$
determine $L = \text{máx}\{x_{j-2}, 0\}$;
$U = \text{mín}\{x_{i+2}, 1\}$;
$a_{ij} = \int_L^U \left[p(x)\phi_i'(x)\phi_j'(x) + q(x)\phi_i(x)\phi_j(x)\right]\,dx$;
si $i \neq j$, entonces determine $a_{ji} = a_{ij}$. (*Puesto que A es simétrica.*)

Paso 7 Si $i \geq 4$ entonces para $j = 0, \ldots, i-4$ determine $a_{ij} = 0$.

Paso 8 Si $i \leq n-3$ entonces para $j = i+4, \ldots, n+1$ determine $a_{ij} = 0$.

Paso 9 Determine $L = \text{máx}\{x_{i-2}, 0\}$;
$U = \text{mín}\{x_{i+2}, 1\}$;
$b_i = \int_L^U f(x)\phi_i(x)\,dx$.

Paso 10 Resolver el sistema lineal $A\mathbf{c} = \mathbf{b}$, donde $A = (a_{ij})$, $\mathbf{b} = (b_0, \ldots, b_{n+1})^t$ y $\mathbf{c} = (c_0, \ldots, c_{n+1})^t$.

Paso 11 Para $i = 0, \ldots, n+1$
SALIDA (c_i).

Paso 12 PARE. (*El procedimiento está completo.*) ■

Ilustración Considere el problema de valor en la frontera

$$-y'' + \pi^2 y = 2\pi^2 \operatorname{sen}(\pi x), \quad \text{para } 0 \leq x \leq 1, \text{con } y(0) = y(1) = 0.$$

En la ilustración después del algoritmo 11.5, hacemos $h = 0.1$ y generamos aproximaciones usando funciones base lineales por tramos. La tabla 11.8 muestra los resultados obtenidos al aplicar los B-splines como se describe en el algoritmo 11.6 con la misma selección de nodos. ■

Tabla 11.8

i	c_i	x_i	$\phi(x_i)$	$y(x_i)$	$\lvert y(x_i) - \phi(x_i)\rvert$
0	$0.50964361 \times 10^{-5}$	0	0.00000000	0.00000000	0.00000000
1	0.20942608	0.1	0.30901644	0.30901699	0.00000055
2	0.39835678	0.2	0.58778549	0.58778525	0.00000024
3	0.54828946	0.3	0.80901687	0.80901699	0.00000012
4	0.64455358	0.4	0.95105667	0.95105652	0.00000015
5	0.67772340	0.5	1.00000002	1.00000000	0.00000020
6	0.64455370	0.6	0.95105713	0.95105652	0.00000061
7	0.54828951	0.7	0.80901773	0.80901699	0.00000074
8	0.39835730	0.8	0.58778690	0.58778525	0.00000165
9	0.20942593	0.9	0.30901810	0.30901699	0.00000111
10	$0.74931285 \times 10^{-5}$	1.0	0.00000000	0.00000000	0.00000000

Recomendamos que las integraciones en los pasos 6 y 9 se realicen en dos partes. Primero, construimos polinomios interpolantes de spline cúbico para p, q y f usando los métodos presentados en la sección 3.5. A continuación, aproximamos los integrandos por productos de splines cúbicos o derivadas de splines cúbicos. Ahora, los integrandos son polinomios por tramos y se pueden integrar exactamente en cada subintervalo y después sumarse. Esto lleva a aproximaciones exactas de las integrales.

Las hipótesis asumidas al principio de esta sección son suficientes para garantizar que

$$\left\{ \int_0^1 |y(x) - \phi(x)|^2 \, dx \right\}^{1/2} = O(h^4), \quad \text{si} \quad 0 \leq x \leq 1.$$

Para una comprobación de este resultado, consulte [Schull], p. 107-108.

Los B-splines también se pueden definir para nodos no espaciados equitativamente, pero los detalles son más complicados. Una presentación de la técnica se puede encontrar en [Schull], p. 73. Otra base que se usa de manera común es el polinomio de Hermite cúbico por tramos. Para una excelente presentación de este método, consulte de nuevo [Schull], p. 24ff.

Boris Grigorievich Galerkin (1871–1945) realizó trabajos fundamentales al aplicar las técnicas de aproximación para resolver problemas de valor en la frontera relacionados con problemas de ingeniería civil. Su artículo inicial sobre análisis de elementos finitos fue publicado en 1915 y su manuscrito fundamental sobre placas elásticas delgadas en 1937.

Otros métodos que reciben considerable atención son Galerkin o métodos de "forma débil". Para el problema de valor en la frontera que hemos considerado,

$$-\frac{d}{dx}\left(p(x)\frac{dy}{dx}\right) + q(x)y = f(x), \quad \text{para } 0 \leq x \leq 1, \text{con } y(0) = 0 \text{ y } y(1) = 0,$$

bajo las suposiciones lis mencionadas al principio de esta sección, los métodos de Galerkin y Rayleigh-Ritz se determinan con la ecuación (11.27). Sin embargo, éste no es el caso de un problema de valor en la frontera arbitrario. Un tratamiento sobre las similitudes y las diferencias en los dos métodos y un análisis de la amplia aplicación del método de Galerkin se puede encontrar en [Schull] y en [SF].

La raíz de la palabra "colocación" proviene del Latín "co-" y "locus", lo cual indica "junto con" y "lugar". Es equivalente a lo que llamamos interpolación.

Otra técnica popular para resolver problemas de valor en la frontera es el **método de colocación**.

Este procedimiento comienza seleccionando un conjunto funciones base $\{\phi_1, \ldots, \phi_N\}$, un conjunto de números $\{x_i, \ldots, x_n\}$ en $[0, 1]$ y requieren que una aproximación

$$\sum_{i=1}^{N} c_i \phi_i(x)$$

satisfaga la ecuación diferencial de cada número x_j, para $1 \leq j \leq n$. Si, además, se requiere que $\phi_i(0) = \phi_i(1) = 0$, para $1 \leq i \leq N$, entonces las condiciones de frontera se satisfacen automáticamente. Se ha prestado mucha atención en la literatura a la selección de números $\{x_j\}$ y las funciones base $\{\phi_i\}$. Una selección común es permitir que ϕ_i sea la función base para las funciones spline relativas a una partición $[0, 1]$ y que los nodos $\{x_j\}$ sean puntos gaussianos o raíces de ciertos polinomios ortogonales, transformados en el subintervalo adecuado.

Una comparación de diferentes métodos de colocación y métodos de diferencia finita se encuentra en [Ru]. La conclusión es que los métodos de colocación que utilizan splines de grado superior compiten con las técnicas de diferencia finita que utilizan extrapolación. Otras referencias para métodos de colocación son [DebS] y [LR].

La sección Conjunto de ejercicios 11.5 está disponible en línea. Encuentre la ruta de acceso en las páginas preliminares.

11.6 Software numérico

La biblioteca IMSL tiene muchas subrutinas para problemas de valor en la frontera. Existen tanto métodos de disparo como de diferencia finita. Los métodos de disparo usan la técnica Runge-Kutta-Verner para resolver los problemas de valor inicial relacionados.

La biblioteca NAG también tiene una multitud de subrutinas para resolver problemas de valor en la frontera. Algunos de estos son un método de disparo que usa el método de valor inicial Runge-Kutta-Merson junto con el método de Newton, un método de diferencia finita con el de Newton para resolver un sistema lineal y uno de diferencia finita lineal con base en la colocación.

Existen subrutinas en el paquete ODE contenido en la biblioteca Netlib para resolver problemas de valor en la frontera de dos puntos no lineales y lineales, respectivamente. Estas rutinas están basadas en métodos de múltiples disparos.

Las secciones Preguntas de análisis, Conceptos clave y Revisión del capítulo están disponibles en línea. Encuentre la ruta de acceso en las páginas preliminares.

CAPÍTULO

12 Soluciones numéricas para ecuaciones diferenciales parciales

Introducción

Un cuerpo es *isotrópico* si la conductividad térmica en cada uno de sus puntos es independiente de la dirección del flujo de calor a través del punto. Suponga que k, c y ρ son funciones de (x, y, z) y representan, respectivamente, la conductividad térmica, el calor específico y la densidad de un cuerpo isotrópico en el punto (x, y, z). Entonces la temperatura $u \equiv u(x, y, z, t)$, en el cuerpo se puede encontrar al resolver la ecuación diferencial parcial

$$\frac{\partial}{\partial x}\left(k\frac{\partial u}{\partial x}\right) + \frac{\partial}{\partial y}\left(k\frac{\partial u}{\partial y}\right) + \frac{\partial}{\partial z}\left(k\frac{\partial u}{\partial z}\right) = c\rho\frac{\partial u}{\partial t}.$$

Cuando k, c y ρ son constantes, a esta ecuación se le conoce como ecuación de calor tridimensional y se expresa como

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2} = \frac{c\rho}{k}\frac{\partial u}{\partial t}.$$

Si la frontera del cuerpo es relativamente simple, la solución de esta ecuación se puede encontrar por medio de series de Fourier.

En muchas situaciones, cuando k, c y ρ no son constantes o cuando la frontera es irregular, la solución de la ecuación diferencial parcial se debe obtener por medio de técnicas de aproximación. En este capítulo se presenta una introducción a estas técnicas.

Ecuaciones elípticas

Generalmente las ecuaciones diferenciales parciales se clasifican de manera similar a las secciones cónicas. La ecuación diferencial parcial que consideraremos en la sección 12.1 involucra $u_{xx}(x, y) + u_{yy}(x, y)$ y es una ecuación **elíptica**. A la ecuación elíptica particular que consideraremos se le conoce como **ecuación de Poisson**:

$$\frac{\partial^2 u}{\partial x^2}(x, y) + \frac{\partial^2 u}{\partial y^2}(x, y) = f(x, y).$$

En esta ecuación, suponemos que f describe la entrada para el problema en una región plana R con frontera S. Las ecuaciones de este tipo surgen en el estudio de diferentes problemas físicos independientes del tiempo, como la distribución de estado estable del calor en una región plana y problemas de estado estable bidimensionales que implican fluidos incompresibles.

Siméon-Denis Poisson (1781–1840) era un estudiante de Laplace y Legendre durante los años napoleónicos en Francia. Más adelante, asumió la cátedra en la École Polytechnique, donde trabajó con ecuaciones diferenciales parciales y ordinarias y, después, en la teoría de la probabilidad de la vida.

Deben imponerse restricciones adicionales para obtener una solución única para la ecuación de Poisson. Por ejemplo, el estudio de la distribución de estado estable de calor en una región plana requiere que $f(x, y) \equiv 0$, lo cual resulta en una simplificación para la **ecuación de Laplace**

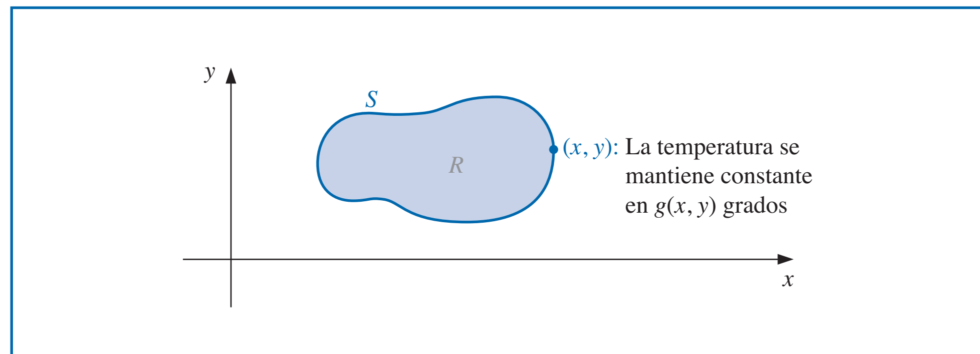
$$\frac{\partial^2 u}{\partial x^2}(x, y) + \frac{\partial^2 u}{\partial y^2}(x, y) = 0.$$

Si la temperatura dentro de la región se determina mediante la distribución de temperatura en la frontera de la región, las restricciones reciben el nombre de **condiciones de frontera de Dirichlet**, dadas por

$$u(x, y) = g(x, y),$$

para todas las (x, y) en S, la frontera de la región R. (Consulte la figura 12.1.)

Figura 12.1



Pierre–Simon Laplace (1749–1827) trabajó en muchas áreas matemáticas, produjo artículos fundamentales sobre probabilidad y física matemática. Publicó su trabajo más importante sobre la teoría del calor durante el periodo de 1817–1820.

Johann Peter Gustav Lejeune Dirichlet (1805–1859) hizo importantes contribuciones a las áreas de la teoría numérica y la convergencia de las series. De hecho, podría ser considerado como el fundador de las series de Fourier ya que, de acuerdo con Riemann, fue el primero en escribir un artículo profundo sobre este tema.

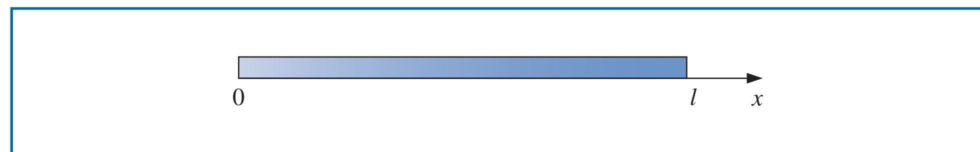
Ecuaciones parabólicas

En la sección 12.2 consideramos la solución numérica del problema que involucra una ecuación diferencial parcial **parabólica** de la forma

$$\frac{\partial u}{\partial t}(x, t) - \alpha^2 \frac{\partial^2 u}{\partial x^2}(x, t) = 0.$$

El problema físico considerado aquí aborda el flujo de calor a lo largo de una varilla (o barra) de longitud l (consulte la figura 12.2) que tiene una temperatura uniforme en cada sección transversal. Esto requiere que la varilla esté perfectamente aislada en su superficie lateral. Se supone que la constante α es independiente de la posición en la varilla. Se determina mediante las propiedades conductoras de calor del material del que ésta se compone

Figura 12.2



Uno de los conjuntos comunes de las restricciones para un problema de flujo de calor de este tipo es especificar la distribución inicial de calor en la varilla,

$$u(x, 0) = f(x),$$

y describir la conducta en sus extremos. Por ejemplo, si los extremos se mantienen a temperaturas constantes U_1 y U_2, las condiciones en la frontera tienen la forma

$$u(0, t) = U_1 \quad \text{y} \quad u(l, t) = U_2,$$

y la distribución de calor se aproxima a la distribución límite de temperatura

$$\lim_{t\to\infty} u(x, t) = U_1 + \frac{U_2 - U_1}{l} x.$$

Si, por el contrario, la varilla se aísla de tal forma que el calor no fluye a través de los extremos, las condiciones en la frontera son

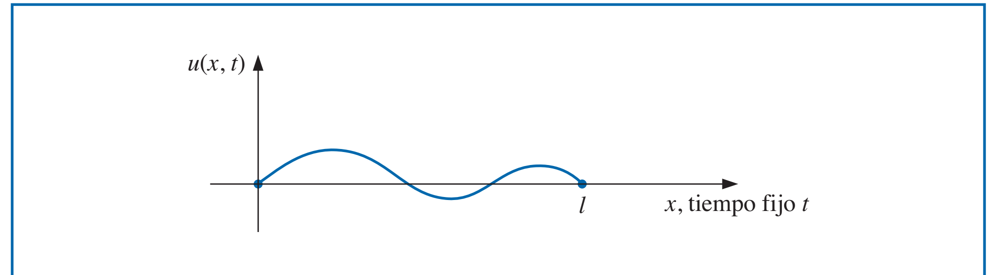
$$\frac{\partial u}{\partial x}(0, t) = 0 \quad \text{y} \quad \frac{\partial u}{\partial x}(l, t) = 0.$$

Entonces, el calor no escapa de la varilla y, en el caso límite, la temperatura en la varilla es constante. La ecuación diferencial parcial parabólica también es importante en el estudio de la difusión de gas; de hecho, es conocida en algunos círculos como **ecuación de difusión**.

Ecuaciones hiperbólicas

El problema estudiado en la sección 12.3 es la **ecuación de onda** y es un ejemplo de una ecuación diferencial parcial **hiperbólica**. Suponga una cuerda elástica de longitud l se estira entre dos soportes en el mismo nivel horizontal (consulte la figura 12.3).

Figura 12.3



Si la cuerda se configura para vibrar en un plano vertical, el desplazamiento vertical $u(x, t)$ de un punto x en el tiempo t satisface la ecuación diferencial parcial

$$\alpha^2 \frac{\partial^2 u}{\partial x^2}(x, t) - \frac{\partial^2 u}{\partial t^2}(x, t) = 0, \quad \text{para } 0 < x < l \quad \text{y} \quad 0 < t,$$

siempre y cuando se ignoren los efectos de amortiguamiento y la amplitud no sea tan grande. Para imponer restricciones sobre este problema, suponemos que la posición inicial y la velocidad de la cuerda están dadas por

$$u(x, 0) = f(x) \quad \text{y} \quad \frac{\partial u}{\partial t}(x, 0) = g(x), \quad \text{para } 0 \leq x \leq l.$$

Si los extremos están fijos, también tenemos $u(0, t) = 0$ y $u(l, t) = 0$.

Otros problemas físicos implican la presencia de la ecuación diferencial parcial hiperbólica en el estudio de vigas que vibran con uno o ambos extremos sujetados y en la transmisión de electricidad en una línea larga donde existe alguna fuga de corriente hacia el piso.

12.1 Ecuaciones diferenciales parciales elípticas

La ecuación diferencial parcial *elíptica* que consideramos es la ecuación de Poisson,

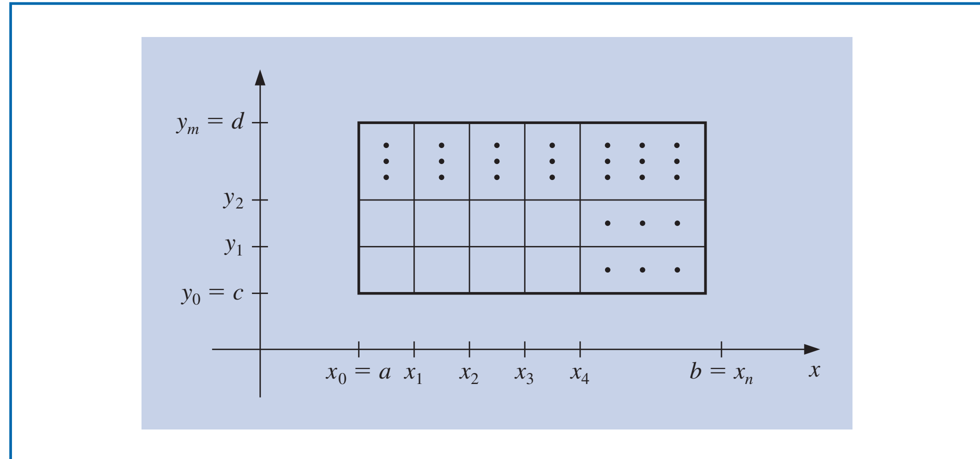
$$\nabla^2 u(x, y) \equiv \frac{\partial^2 u}{\partial x^2}(x, y) + \frac{\partial^2 u}{\partial y^2}(x, y) = f(x, y), \tag{12.1}$$

en $R = \{ (x, y) \mid a < x < b, c < y < d \}$, con $u(x, y) = g(x, y)$ para $(x, y) \in S$, donde S denota la frontera de R. Si f y g son continuas en sus dominios, entonces existe una única solución para esta ecuación.

Selección de una cuadrícula

El método utilizado es una adaptación bidimensional del método de diferencias finitas para problemas lineales de valor en la frontera, analizados en la sección 11.3. El primer paso es seleccionar enteros n y m para definir los tamaños de paso $h = (b - a)/n$ y $k = (d - c)/m$. La división del intervalo $[a, b]$ en n partes iguales, de ancho h y el intervalo $[c, d]$ en m partes iguales, de ancho k (consulte la figura 12.4).

Figura 12.4



Coloque una cuadrícula sobre el rectángulo **R** al trazar líneas verticales y horizontales a través de los puntos con coordenadas (x_i, y_j), donde

$$x_i = a + ih, \quad \text{para cada } i = 0, 1, \ldots, n, \quad \text{y} \quad y_j = c + jk, \quad \text{para cada } j = 0, 1, \ldots, m.$$

Las líneas $x = x_i$ y $y = y_j$ son **líneas de cuadrícula** y sus intersecciones son los **puntos de malla** de la cuadrícula. Para cada punto de malla en el interior de la cuadrícula (x_i, y_j), para $i = 1, 2, \ldots, n - 1$ y $j = 1, 2, \ldots, m - 1$, podemos usar la serie de Taylor en la variable x alrededor de x_i para generar la fórmula de diferencias centradas

$$\frac{\partial^2 u}{\partial x^2}(x_i, y_j) = \frac{u(x_{i+1}, y_j) - 2u(x_i, y_j) + u(x_{i-1}, y_j)}{h^2} - \frac{h^2}{12}\frac{\partial^4 u}{\partial x^4}(\xi_i, y_j), \tag{12.2}$$

donde $\xi_i \in (x_{i-1}, x_{i+1})$. También podemos usar la serie de Taylor en la variable y alrededor de y_j para generar la fórmula de diferencias centradas

$$\frac{\partial^2 u}{\partial y^2}(x_i, y_j) = \frac{u(x_i, y_{j+1}) - 2u(x_i, y_j) + u(x_i, y_{j-1})}{k^2} - \frac{k^2}{12}\frac{\partial^4 u}{\partial y^4}(x_i, \eta_j), \tag{12.3}$$

donde $\eta_j \in (y_{j-1}, y_{j+1})$.

Por medio de estas fórmulas en la ecuación (12.1) podemos expresar la ecuación de Poisson en los puntos (x_i, y_j) como

$$\frac{u(x_{i+1}, y_j) - 2u(x_i, y_j) + u(x_{i-1}, y_j)}{h^2} + \frac{u(x_i, y_{j+1}) - 2u(x_i, y_j) + u(x_i, y_{j-1})}{k^2}$$
$$= f(x_i, y_j) + \frac{h^2}{12}\frac{\partial^4 u}{\partial x^4}(\xi_i, y_j) + \frac{k^2}{12}\frac{\partial^4 u}{\partial y^4}(x_i, \eta_j),$$

para cada $i = 1, 2, \ldots, n-1$ y $j = 1, 2, \ldots, m-1$. Las condiciones de frontera son

$$u(x_0, y_j) = g(x_0, y_j) \quad \text{y} \quad u(x_n, y_j) = g(x_n, y_j), \quad \text{para cada } j = 0, 1, \ldots, m;$$
$$u(x_i, y_0) = g(x_i, y_0) \quad \text{y} \quad u(x_i, y_m) = g(x_i, y_m), \quad \text{para cada } i = 1, 2, \ldots, n-1.$$

Método de diferencias finitas

En la forma de ecuación de diferencia, esto resulta en el **método de diferencias finitas**:

$$2\left[\left(\frac{h}{k}\right)^2 + 1\right] w_{ij} - (w_{i+1,j} + w_{i-1,j}) - \left(\frac{h}{k}\right)^2 (w_{i,j+1} + w_{i,j-1}) = -h^2 f(x_i, y_j), \tag{12.4}$$

para cada $i = 1, 2, \ldots, n-1$ y $j = 1, 2, \ldots, m-1$, y

$$w_{0j} = g(x_0, y_j) \quad \text{y} \quad w_{nj} = g(x_n, y_j), \quad \text{para cada } j = 0, 1, \ldots, m; \tag{12.5}$$
$$w_{i0} = g(x_i, y_0) \quad \text{y} \quad w_{im} = g(x_i, y_m), \quad \text{para cada } i = 1, 2, \ldots, n-1;$$

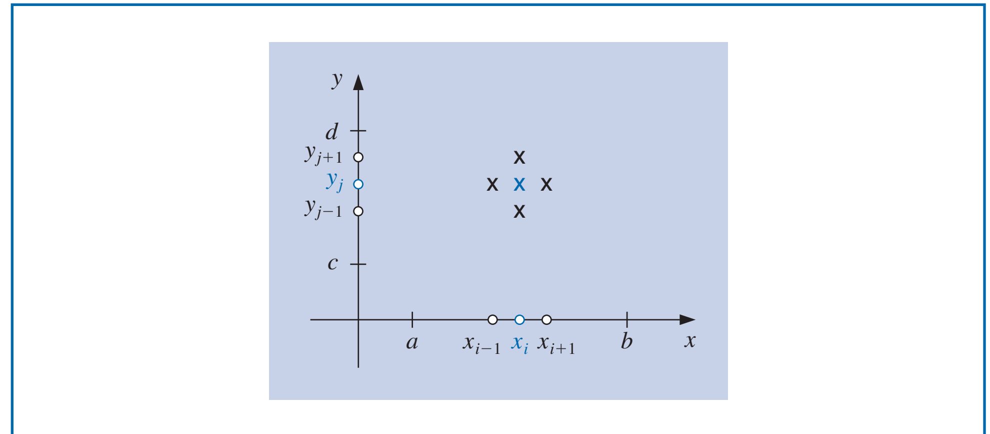
donde w_{ij} aproxima $u(x_i, y_j)$. Este método tiene error de truncamiento local de orden $O(h^2 + k^2)$.

La ecuación común en (12.4) implica aproximaciones para $u(x, y)$ en los puntos

$$(x_{i-1}, y_j), \quad (x_i, y_j), \quad (x_{i+1}, y_j), \quad (x_i, y_{j-1}), \quad \text{y} \quad (x_i, y_{j+1}).$$

Reproducir la parte de la cuadrícula en la que se localizan estos puntos (consulte la figura 12.5) muestra que cada ecuación implica aproximaciones en la región en forma de estrella alrededor de la x en (x_i, y_j).

Figura 12.5



Usamos la información a partir de las condiciones de frontera (12.5) siempre que sea apropiado en el sistema dado por la ecuación (12.4), es decir, en todos los pun-

tos (x_i, y_j) adyacentes a un punto de malla en la frontera. Esto produce un sistema lineal $(n-1)(m-1) \times (n-1)(m-1)$ con las incógnitas como aproximaciones $w_{i\,j}$ para $u(x_i, y_j)$ en los puntos de malla interiores.

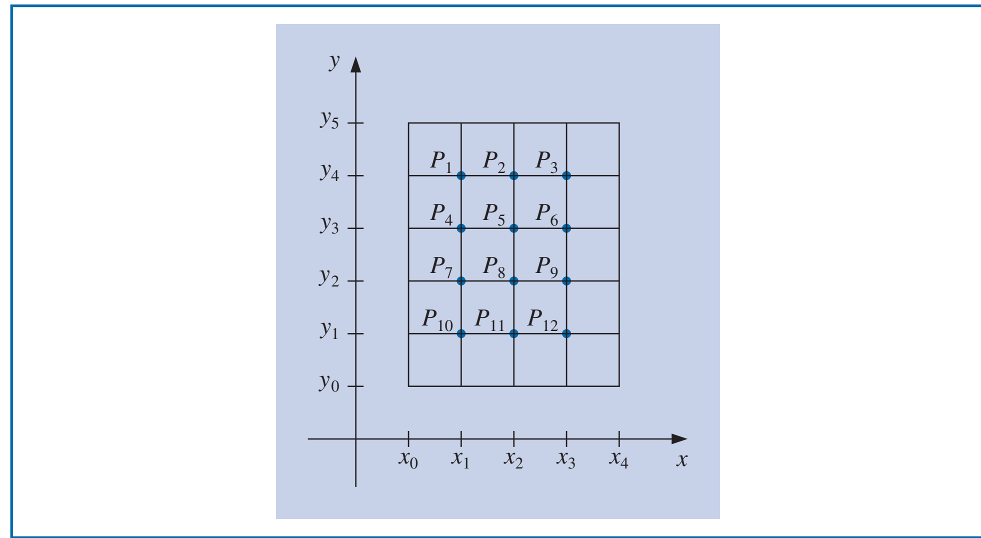
El sistema lineal que contiene estas incógnitas se expresa, para los cálculos de matriz, de forma más eficiente si se introduce el reetiquetado de los puntos de malla interiores. Un etiquetado recomendado de estos puntos (consulte [Var1], p. 210) es hacer

$$P_l = (x_i, y_j) \quad \text{y} \quad w_l = w_{i,j},$$

donde $l = i + (m-1-j)(n-1)$, para cada $i = 1, 2, \ldots, n-1$ y $j = 1, 2, \ldots, m-1$. Esto etiqueta los puntos de malla de forma consecutiva desde la izquierda hasta la derecha y desde la parte superior hasta la parte inferior. Etiquetar los puntos de esta forma garantiza que el sistema necesario para determinar $w_{i,j}$ es una matriz con un ancho de banda como máximo $2n-1$.

Por ejemplo, con $n = 4$ y $m = 5$, el reetiquetado resulta en una cuadrícula cuyos puntos se muestran en la figura 12.6.

Figura 12.6



Ejemplo 1 Determine la distribución de calor en estado estable en una placa de metal cuadrada y delgada con dimensiones de 0.5 m por 0.5 m usando $n = m = 4$. Dos fronteras adyacentes se mantienen a 0°C y el calor en las otras fronteras se incrementa de manera lineal desde 0ºC en una esquina hasta 100°C donde se unen los lados.

Solución Coloque los lados con las condiciones de frontera cero a lo largo de los ejes x y y. Entonces, el problema se expresa como

$$\frac{\partial^2 u}{\partial x^2}(x, y) + \frac{\partial^2 u}{\partial y^2}(x, y) = 0,$$

para (x, y) en el conjunto $R = \{(x, y) \mid 0 < x < 0.5,\ 0 < y < 0.5\}$. Las condiciones de frontera son

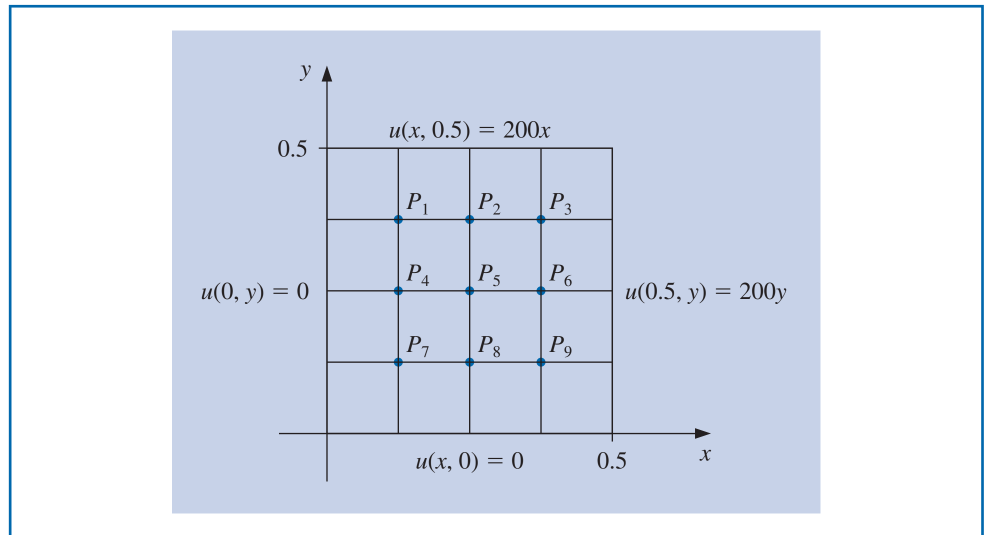
$$u(0, y) = 0, \quad u(x, 0) = 0, \quad u(x, 0.5) = 200x, \text{ y } u(0.5, y) = 200y.$$

Si $n = m = 4$, el problema tiene la cuadrícula que se muestra en la figura 12.7 y la ecuación de diferencias (12.4) es

$$4w_{i,j} - w_{i+1,j} - w_{i-1,j} - w_{i,j-1} - w_{i,j+1} = 0,$$

para cada $i = 1, 2, 3$ y $j = 1, 2, 3$.

Figura 12.7



Expresar esto en términos de puntos de cuadrícula interior reetiquetados $w_i = u(P_i)$ implica que las ecuaciones en los puntos P_i son

$$
\begin{aligned}
P_1 :&\quad 4w_1 - w_2 - w_4 = w_{0,3} + w_{1,4},\\
P_2 :&\quad 4w_2 - w_3 - w_1 - w_5 = w_{2,4},\\
P_3 :&\quad 4w_3 - w_2 - w_6 = w_{4,3} + w_{3,4},\\
P_4 :&\quad 4w_4 - w_5 - w_1 - w_7 = w_{0,2},\\
P_5 :&\quad 4w_5 - w_6 - w_4 - w_2 - w_8 = 0,\\
P_6 :&\quad 4w_6 - w_5 - w_3 - w_9 = w_{4,2},\\
P_7 :&\quad 4w_7 - w_8 - w_4 = w_{0,1} + w_{1,0},\\
P_8 :&\quad 4w_8 - w_9 - w_7 - w_5 = w_{2,0},\\
P_9 :&\quad 4w_9 - w_8 - w_6 = w_{3,0} + w_{4,1},
\end{aligned}
$$

donde los lados derechos de las ecuaciones se obtienen a partir de las condiciones de frontera.

De hecho, las condiciones de frontera implican que

$$w_{1,0} = w_{2,0} = w_{3,0} = w_{0,1} = w_{0,2} = w_{0,3} = 0,$$

$$w_{1,4} = w_{4,1} = 25, \quad w_{2,4} = w_{4,2} = 50, \quad \text{y} \quad w_{3,4} = w_{4,3} = 75.$$

Por lo que el sistema lineal asociado con este problema tiene la forma

$$
\begin{bmatrix}
4 & -1 & 0 & -1 & 0 & 0 & 0 & 0 & 0\\
-1 & 4 & -1 & 0 & -1 & 0 & 0 & 0 & 0\\
0 & -1 & 4 & 0 & 0 & -1 & 0 & 0 & 0\\
-1 & 0 & 0 & 4 & -1 & 0 & -1 & 0 & 0\\
0 & -1 & 0 & -1 & 4 & -1 & 0 & -1 & 0\\
0 & 0 & -1 & 0 & -1 & 4 & 0 & 0 & -1\\
0 & 0 & 0 & -1 & 0 & 0 & 4 & -1 & 0\\
0 & 0 & 0 & 0 & -1 & 0 & -1 & 4 & -1\\
0 & 0 & 0 & 0 & 0 & -1 & 0 & -1 & 4
\end{bmatrix}
\begin{bmatrix} w_1\\ w_2\\ w_3\\ w_4\\ w_5\\ w_6\\ w_7\\ w_8\\ w_9 \end{bmatrix}
=
\begin{bmatrix} 25\\ 50\\ 150\\ 0\\ 0\\ 50\\ 0\\ 0\\ 25 \end{bmatrix}.
$$

Los valores de $w_1, w_2, \ldots, w_9$, encontrados al aplicar el método Gauss-Seidel a esta matriz se establecen en la tabla 12.1.

Tabla 12.1

i	w_i
1	18.75
2	37.50
3	56.25
4	12.50
5	25.00
6	37.50
7	6.25
8	12.50
9	18.75

Estas respuestas son exactas porque la verdadera solución, $u(x, y) = 400xy$, tiene

$$\frac{\partial^4 u}{\partial x^4} = \frac{\partial^4 u}{\partial y^4} \equiv 0,$$

y el error de truncamiento es cero en cada paso. ■

El problema considerado en el ejemplo 1 tiene el mismo tamaño de malla, 0.125, en cada eje y requiere resolver solamente un sistema lineal 9×9. Esto simplifica la situación y no introduce los problemas computacionales presentes cuando el sistema es más grande. El algoritmo 12.1 usa el método iterativo de Gauss-Seidel para resolver el sistema lineal que resulta y permite tamaños de malla desiguales en los ejes.

ALGORITMO 12.1

Diferencia finita de la ecuación de Poisson

Para aproximar la solución de la ecuación de Poisson

$$\frac{\partial^2 u}{\partial x^2}(x, y) + \frac{\partial^2 u}{\partial y^2}(x, y) = f(x, y), \quad a \le x \le b, \quad c \le y \le d,$$

sujeta a las condiciones de frontera

$$u(x, y) = g(x, y) \quad \text{si } x = a \text{ o } x = b \quad \text{y} \quad c \le y \le d$$

y

$$u(x, y) = g(x, y) \quad \text{si } y = c \text{ o } y = d \quad \text{y} \quad a \le x \le b:$$

ENTRADA extremos a, b, c, d; enteros $m \ge 3, n \ge 3$; tolerancia TOL; número máximo de iteraciones N.

SALIDA aproximaciones $w_{i,j}$ para $u(x_i, y_j)$ para cada $i = 1, \ldots, n-1$ y para cada $j = 1, \ldots, m-1$ o un mensaje que indica que se excedió el número máximo de iteraciones.

Paso 1 Determine $h = (b - a)/n$;
$k = (d - c)/m$.

Paso 2 Para $i = 1, \ldots, n-1$ determine $x_i = a + ih$. (*Los pasos* 2 *y* 3 *construyen puntos de malla.*)

Paso 3 Para $j = 1, \ldots, m-1$ determine $y_j = c + jk$.

Paso 4 Para $i = 1, \ldots, n-1$
para $j = 1, \ldots, m-1$ determine $w_{i,j} = 0$.

Paso 5 Determine $\lambda = h^2/k^2$;
$\mu = 2(1 + \lambda)$;
$l = 1$.

Paso 6 Mientras $l \le N$ haga los pasos 7–20. (*Los pasos 7–20 realizan iteraciones de Gauss-Seidel.*)

Paso 7 Determine $z = (-h^2 f(x_1, y_{m-1}) + g(a, y_{m-1}) + \lambda g(x_1, d)$
$+ \lambda w_{1,m-2} + w_{2,m-1})/\mu$;
$NORM = |z - w_{1,m-1}|$;
$w_{1,m-1} = z$.

Paso 8 Para $i = 2, \ldots, n-2$

determine $z = (-h^2 f(x_i, y_{m-1}) + \lambda g(x_i, d) + w_{i-1,m-1}$
$+w_{i+1,m-1} + \lambda w_{i,m-2})/\mu$;
si $|w_{i,m-1} - z| > NORM$ entonces establezca $NORM = |w_{i,m-1} - z|$;
determine $w_{i,m-1} = z$.

Paso 9 Determine $z = (-h^2 f(x_{n-1}, y_{m-1}) + g(b, y_{m-1}) + \lambda g(x_{n-1}, d)$
$+w_{n-2,m-1} + \lambda w_{n-1,m-2})/\mu$;
si $|w_{n-1,m-1} - z| > NORM$ entonces determine $NORM = |w_{n-1,m-1} - z|$;
determine $w_{n-1,m-1} = z$.

Paso 10 Para $j = m - 2, \ldots, 2$ haga los pasos 11, 12, y 13.

Paso 11 Determine $z = (-h^2 f(x_1, y_j) + g(a, y_j) + \lambda w_{1,j+1}$
$+\lambda w_{1,j-1} + w_{2,j})/\mu$;
si $|w_{1,j} - z| > NORM$ entonces determine $NORM = |w_{1,j} - z|$;
determine $w_{1,j} = z$.

Paso 12 Para $i = 2, \ldots, n-2$
determine $z = (-h^2 f(x_i, y_j) + w_{i-1,j} + \lambda w_{i,j+1}$
$+w_{i+1,j} + \lambda w_{i,j-1})/\mu$;
si $|w_{i,j} - z| > NORM$ entonces determine $NORM = |w_{i,j} - z|$;
determine $w_{i,j} = z$.

Paso 13 Determine $z = (-h^2 f(x_{n-1}, y_j) + g(b, y_j) + w_{n-2,j}$
$+\lambda w_{n-1,j+1} + \lambda w_{n-1,j-1})/\mu$;
si $|w_{n-1,j} - z| > NORM$ entonces determine $NORM = |w_{n-1,j} - z|$;
determine $w_{n-1,j} = z$.

Paso 14 Determine $z = (-h^2 f(x_1, y_1) + g(a, y_1) + \lambda g(x_1, c) + \lambda w_{1,2} + w_{2,1})/\mu$;
si $|w_{1,1} - z| > NORM$ entonces determine $NORM = |w_{1,1} - z|$;
determine $w_{1,1} = z$.

Paso 15 Para $i = 2, \ldots, n-2$
determine $z = (-h^2 f(x_i, y_1) + \lambda g(x_i, c) + w_{i-1,1} + \lambda w_{i,2} + w_{i+1,1})/\mu$;
si $|w_{i,1} - z| > NORM$ entonces determine $NORM = |w_{i,1} - z|$;
determine $w_{i,1} = z$.

Paso 16 Determine $z = (-h^2 f(x_{n-1}, y_1) + g(b, y_1) + \lambda g(x_{n-1}, c)$
$+w_{n-2,1} + \lambda w_{n-1,2})/\mu$;
si $|w_{n-1,1} - z| > NORM$ entonces determine $NORM = |w_{n-1,1} - z|$;
determine $w_{n-1,1} = z$.

Paso 17 Si $NORM \leq TOL$ entonces hacer los pasos 18 y 19.

Paso 18 Para $i = 1, \ldots, n-1$
para $j = 1, \ldots, m-1$ SALIDA $(x_i, y_j, w_{i,j})$.

Paso 19 PARE. (*El procedimiento fue exitoso.*)

Paso 20 Determine $l = l + 1$.

Paso 21 SALIDA ('Número máximo de iteracciones excedido');
(*El procedimiento no fue exitoso.*)
PARE. ■

A pesar de que el procedimiento iterativo de Gauss-Seidel se incorpora en el algoritmo 12.1 para simplicidad, es aconsejable usar una técnica directa como eliminación gaussiana cuando el sistema es pequeño, del orden de 100 o menos, porque el carácter definida positiva

garantiza estabilidad respecto a errores de redondeo. Especialmente, una generalización del algoritmo de factorización de Crout 6.7 (consulte [Var1], p. 221) es eficiente para resolver este sistema porque la matriz es de la forma tridiagonal simétrica por bloques

$$\begin{bmatrix} A_1 & C_1 & 0 & \cdots & 0 \\ C_1 & A_2 & C_2 & \ddots & \vdots \\ 0 & C_2 & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & C_{m-1} \\ 0 & \cdots & 0 & C_{m-1} & A_{m-1} \end{bmatrix},$$

con bloques cuadrados de tamaño $(n-1) \times (n-1)$.

Selección del método iterativo

Para sistemas grandes se debería usar un método iterativo, específicamente, el método SOR analizado en el algoritmo 7.3. La selección de ω que es óptima en esta situación proviene del hecho de que cuando A se descompone en su diagonal D y partes triangular superior y triangular inferior U y L,

$$A = D - L - U,$$

y B es la matriz para el método de Jacobi,

$$B = D^{-1}(L + U),$$

entonces, el radio espectral de B es (consulte [Var1])

$$\rho(B) = \frac{1}{2}\left[\cos\left(\frac{\pi}{m}\right) + \cos\left(\frac{\pi}{n}\right)\right].$$

El valor de ω que se va a usar es, por consiguiente,

$$\omega = \frac{2}{1 + \sqrt{1 - [\rho(B)]^2}} = \frac{4}{2 + \sqrt{4 - \left[\cos\left(\frac{\pi}{m}\right) + \cos\left(\frac{\pi}{n}\right)\right]^2}}.$$

Una técnica de bloque se puede incorporar al algoritmo para convergencia más rápida del procedimiento SOR. Para una presentación de esta técnica, consulte [Var1], p. 219–223.

Ejemplo 2 Use el método de diferencias finitas de Poisson con $n = 6$, $m = 5$, y una tolerancia de 10^{-10} para aproximar la solución de

$$\frac{\partial^2 u}{\partial x^2}(x, y) + \frac{\partial^2 u}{\partial y^2}(x, y) = xe^y, \quad 0 < x < 2, \quad 0 < y < 1,$$

con las condiciones de frontera

$$u(0, y) = 0, \quad u(2, y) = 2e^y, \quad 0 \le y \le 1,$$
$$u(x, 0) = x, \quad u(x, 1) = ex, \quad 0 \le x \le 2,$$

y compare los resultados con la solución exacta $u(x, y) = xe^y$.

Solución Usando el algoritmo 12.1 con un número máximo de iteraciones establecidas en $N = 100$ proporciona los resultados en la tabla 12.2. El criterio de detener el método Gauss–Seidel en el paso 17 requiere que

$$\left| w_{ij}^{(l)} - w_{ij}^{(l-1)} \right| \leq 10^{-10},$$

para cada $i = 1, \ldots, 5$ y $j = 1, \ldots, 4$. La solución de la ecuación de diferencias se obtuvo con exactitud y el procedimiento se detuvo en $l = 61$. Los resultados, junto con los valores correctos, se presentan en la tabla 12.2. ■

Tabla 12.2

i	j	x_i	y_j	$w_{i,j}^{(61)}$	$u(x_i, y_j)$	$\left\|u(x_i, y_j) - w_{i,j}^{(61)}\right\|$
1	1	0.3333	0.2000	0.40726	0.40713	1.30×10^{-4}
1	2	0.3333	0.4000	0.49748	0.49727	2.08×10^{-4}
1	3	0.3333	0.6000	0.60760	0.60737	2.23×10^{-4}
1	4	0.3333	0.8000	0.74201	0.74185	1.60×10^{-4}
2	1	0.6667	0.2000	0.81452	0.81427	2.55×10^{-4}
2	2	0.6667	0.4000	0.99496	0.99455	4.08×10^{-4}
2	3	0.6667	0.6000	1.2152	1.2147	4.37×10^{-4}
2	4	0.6667	0.8000	1.4840	1.4837	3.15×10^{-4}
3	1	1.0000	0.2000	1.2218	1.2214	3.64×10^{-4}
3	2	1.0000	0.4000	1.4924	1.4918	5.80×10^{-4}
3	3	1.0000	0.6000	1.8227	1.8221	6.24×10^{-4}
3	4	1.0000	0.8000	2.2260	2.2255	4.51×10^{-4}
4	1	1.3333	0.2000	1.6290	1.6285	4.27×10^{-4}
4	2	1.3333	0.4000	1.9898	1.9891	6.79×10^{-4}
4	3	1.3333	0.6000	2.4302	2.4295	7.35×10^{-4}
4	4	1.3333	0.8000	2.9679	2.9674	5.40×10^{-4}
5	1	1.6667	0.2000	2.0360	2.0357	3.71×10^{-4}
5	2	1.6667	0.4000	2.4870	2.4864	5.84×10^{-4}
5	3	1.6667	0.6000	3.0375	3.0369	6.41×10^{-4}
5	4	1.6667	0.8000	3.7097	3.7092	4.89×10^{-4}

La sección Conjunto de ejercicios 12.1 está disponible en línea. Encuentre la ruta de acceso en las páginas preliminares.



12.2 Ecuaciones diferenciales parciales parabólicas

La ecuación diferencial parcial *parabólica* que consideramos es la ecuación de calor, o difusión

$$\frac{\partial u}{\partial t}(x, t) = \alpha^2 \frac{\partial^2 u}{\partial x^2}(x, t), \quad 0 < x < l, \quad t > 0, \tag{12.6}$$

sujeta a las condiciones

$$u(0, t) = u(l, t) = 0, \quad t > 0, \quad \text{y} \quad u(x, 0) = f(x), \quad 0 \leq x \leq l.$$

El enfoque que usamos para aproximar la solución de este problema implica diferencias finitas y es similar al método que se usó en la sección 12.1.

En primer lugar, seleccione un entero $m > 0$ y defina el tamaño de longitud de paso del eje x $h = l/m$. A continuación, seleccione un tamaño de longitud de paso de tiempo k. Los puntos de cuadrícula para esta situación son (x_i, t_j), donde $x_i = ih$, para $i = 0, 1, \ldots, m$, y $t_j = jk$, para $j = 0, 1, \ldots$.

Método de diferencias progresivas

Nosotros obtenemos el método de diferencias mediante la serie de Taylor en t para formar el cociente de diferencias

$$\frac{\partial u}{\partial t}(x_i, t_j) = \frac{u(x_i, t_j + k) - u(x_i, t_j)}{k} - \frac{k}{2}\frac{\partial^2 u}{\partial t^2}(x_i, \mu_j), \tag{12.7}$$

para alguna $\mu_j \in (t_j, t_{j+1})$, y la serie de Taylor en x para formar el cociente de diferencias

$$\frac{\partial^2 u}{\partial x^2}(x_i, t_j) = \frac{u(x_i + h, t_j) - 2u(x_i, t_j) + u(x_i - h, t_j)}{h^2} - \frac{h^2}{12}\frac{\partial^4 u}{\partial x^4}(\xi_i, t_j), \tag{12.8}$$

donde $\xi_i \in (x_{i-1}, x_{i+1})$.

La ecuación diferencial parcial parabólica (12.6) implica que para los puntos en el interior de la malla (x_i, t_j), para cada $i = 1, 2, \ldots, m-1$ y $j = 1, 2, \ldots,$ tenemos

$$\frac{\partial u}{\partial t}(x_i, t_j) - \alpha^2 \frac{\partial^2 u}{\partial x^2}(x_i, t_j) = 0,$$

así que el método usa los cocientes de diferencias (12.7) y (12.8) es

$$\frac{w_{i,j+1} - w_{ij}}{k} - \alpha^2 \frac{w_{i+1,j} - 2w_{ij} + w_{i-1,j}}{h^2} = 0, \tag{12.9}$$

donde w_{ij} aproxima a $u(x_i, t_j)$

El error de truncamiento local para esta ecuación de diferencias es

$$\tau_{ij} = \frac{k}{2}\frac{\partial^2 u}{\partial t^2}(x_i, \mu_j) - \alpha^2 \frac{h^2}{12}\frac{\partial^4 u}{\partial x^4}(\xi_i, t_j). \tag{12.10}$$

Resolviendo la ecuación (12.9) para $w_{i,j+1}$ obtenemos

$$w_{i,j+1} = \left(1 - \frac{2\alpha^2 k}{h^2}\right) w_{ij} + \alpha^2 \frac{k}{h^2}(w_{i+1,j} + w_{i-1,j}), \tag{12.11}$$

para cada $i = 1, 2, \ldots, m-1$ y $j = 1, 2, \ldots$.

Así, obtenemos

$$w_{0,0} = f(x_0), \quad w_{1,0} = f(x_1), \quad \ldots w_{m,0} = f(x_m).$$

Luego generamos la siguiente t-fila por

$$\begin{aligned}
w_{0,1} &= u(0, t_1) = 0;\\
w_{1,1} &= \left(1 - \frac{2\alpha^2 k}{h^2}\right) w_{1,0} + \alpha^2 \frac{k}{h^2}(w_{2,0} + w_{0,0});\\
w_{2,1} &= \left(1 - \frac{2\alpha^2 k}{h^2}\right) w_{2,0} + \alpha^2 \frac{k}{h^2}(w_{3,0} + w_{1,0});\\
&\vdots\\
w_{m-1,1} &= \left(1 - \frac{2\alpha^2 k}{h^2}\right) w_{m-1,0} + \alpha^2 \frac{k}{h^2}(w_{m,0} + w_{m-2,0});\\
w_{m,1} &= u(m, t_1) = 0.
\end{aligned}$$

Ahora podemos usar los valores $w_{i,1}$ para generar todos los valores de la forma $w_{i,2}$.

La naturaleza explícita del método de diferencias implica que la matriz $(m-1) \times (m-1)$ asociada con este sistema puede escribirse en forma tridiagonal

$$A = \begin{bmatrix} (1-2\lambda) & \lambda & 0 & \cdots & 0 \\ \lambda & (1-2\lambda) & \lambda & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & \ddots & \lambda \\ 0 & \cdots & 0 & \lambda & (1-2\lambda) \end{bmatrix},$$

donde $\lambda = \alpha^2(k/h^2)$. Si hacemos

$$\mathbf{w}^{(0)} = (f(x_1), f(x_2), \ldots, f(x_{m-1}))^t$$

y

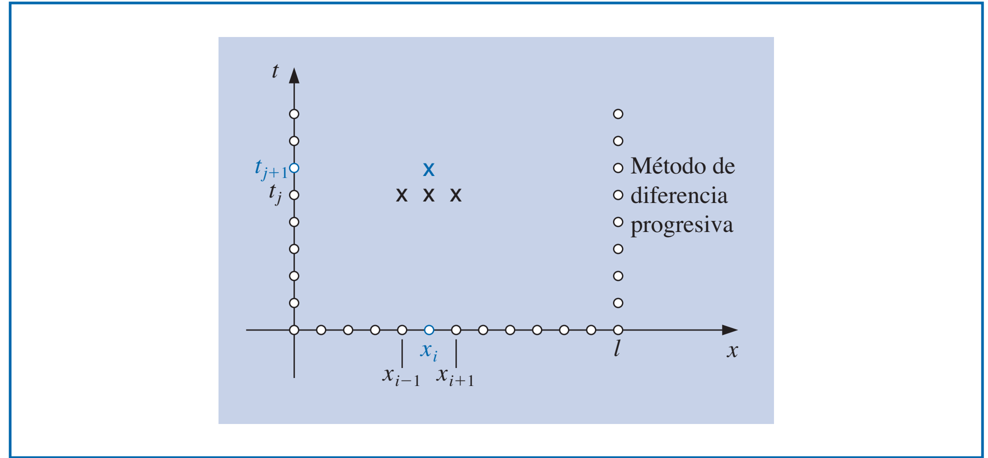
$$\mathbf{w}^{(j)} = (w_{1j}, w_{2j}, \ldots, w_{m-1,j})^t \quad \text{para cada } j = 1, 2, \ldots,$$

entonces, la solución aproximada está dada por

$$\mathbf{w}^{(j)} = A\mathbf{w}^{(j-1)}, \quad \text{para cada } j = 1, 2, \ldots,$$

por lo que $\mathbf{w}^{(j)}$ se obtiene a partir de $\mathbf{w}^{(j-1)}$ mediante una simple multiplicación de matriz. A esto se le conoce como el **método de diferencias progresivas** y la aproximación en el punto azul mostrado en la figura 12.8 usa información de otros puntos marcados en esa figura. Si la solución de la ecuación diferencial parcial tiene cuatro derivadas parciales en x y dos en t, entonces la ecuación (12.10) implica que el método es de orden $O(k + h^2)$.

Figura 12.8



Ejemplo 1 Use los tamaños de paso **a)** $h = 0.1$ y $k = 0.0005$ y **b)** $h = 0.1$ y $k = 0.01$ para aproximar la solución de la ecuación de calor

$$\frac{\partial u}{\partial t}(x,t) - \frac{\partial^2 u}{\partial x^2}(x,t) = 0, \quad 0 < x < 1, \quad 0 \leq t,$$

con condiciones de frontera

$$u(0,t) = u(1,t) = 0, \quad 0 < t,$$

y condiciones iniciales

$$u(x, 0) = \operatorname{sen}(\pi x), \quad 0 \leq x \leq 1.$$

Compare los resultados en $t = 0.5$ con la solución exacta

$$u(x, t) = e^{-\pi^2 t} \operatorname{sen}(\pi x).$$

Solución **a)** El método de diferencias progresivas con $h = 0.1$, $k = 0.0005$, y $\lambda = (1)^2(0.0005/(0.1)^2) = 0.05$ da los resultados en la tercera columna de la tabla 12.3. Como se puede observar a partir de la cuarta columna, estos resultados son bastante exactos.

b) El método de diferencias progresivas con $h = 0.1$, $k = 0.01$ y $\lambda = (1)^2(0.01/(0.1)^2) = 1$ da los resultados en la quinta columna de la tabla 12.3. Como se puede observar a partir de la sexta columna, estos resultados son inútiles. ■

Tabla 12.3

x_i	$u(x_i, 0.5)$	$w_{i,1000}$ $k = 0.0005$	$\lvert u(x_i, 0.5) - w_{i,1000}\rvert$	$w_{i,50}$ $k = 0.01$	$\lvert u(x_i, 0.5) - w_{i,50}\rvert$
0.0	0	0		0	
0.1	0.00222241	0.00228652	6.411×10^{-5}	8.19876×10^{7}	8.199×10^{7}
0.2	0.00422728	0.00434922	1.219×10^{-4}	-1.55719×10^{8}	1.557×10^{8}
0.3	0.00581836	0.00598619	1.678×10^{-4}	2.13833×10^{8}	2.138×10^{8}
0.4	0.00683989	0.00703719	1.973×10^{-4}	-2.50642×10^{8}	2.506×10^{8}
0.5	0.00719188	0.00739934	2.075×10^{-4}	2.62685×10^{8}	2.627×10^{8}
0.6	0.00683989	0.00703719	1.973×10^{-4}	-2.49015×10^{8}	2.490×10^{8}
0.7	0.00581836	0.00598619	1.678×10^{-4}	2.11200×10^{8}	2.112×10^{8}
0.8	0.00422728	0.00434922	1.219×10^{-4}	-1.53086×10^{8}	1.531×10^{8}
0.9	0.00222241	0.00228652	6.511×10^{-5}	8.03604×10^{7}	8.036×10^{7}
1.0	0	0		0	

Consideraciones de estabilidad

En el ejemplo 1 se espera un error de truncamiento de orden $O(k + h^2)$ A pesar de que esto se obtiene con $h = 0.1$ y $k = 0.0005$, sin duda alguna no se obtiene cuando $h = 0.1$ y $k = 0.01$. Para explicar la dificultad, necesitamos observar la estabilidad del método de diferencias progresivas.

Suponga que se comete un error $\mathbf{e}^{(0)} = \left(e_1^{(0)}, e_2^{(0)}, \ldots, e_{m-1}^{(0)}\right)^t$ al representar los datos iniciales

$$\mathbf{w}^{(0)} = \left(f(x_1), f(x_2), \ldots, f(x_{m-1})\right)^t$$

(o en cualquier paso particular, la selección del paso inicial es simplemente por conveniencia). Un error de $A\mathbf{e}^{(0)}$ se propaga en $\mathbf{w}^{(1)}$ porque

$$\mathbf{w}^{(1)} = A\left(\mathbf{w}^{(0)} + \mathbf{e}^{(0)}\right) = A\mathbf{w}^{(0)} + A\mathbf{e}^{(0)}.$$

Este proceso continúa en el enésimo paso, el error en $\mathbf{w}^{(n)}$ debido a $\mathbf{e}^{(0)}$ es $A^n\mathbf{e}^{(0)}$. Por consiguiente, el método es estable precisamente cuando estos errores no crecen conforme n se incrementa. Pero esto no es verdadero si y sólo si para cualquier error inicial $\mathbf{e}^{(0)}$, tenemos $\left\|A^n\mathbf{e}^{(0)}\right\| \leq \left\|\mathbf{e}^{(0)}\right\|$ para todas las n. Por lo tanto, debemos tener $\|A^n\| \leq 1$, una condición que, mediante el teorema 7.15 en la página 332, requiere que $\rho(A^n) = (\rho(A))^n \leq 1$. El método de diferencias progresivas es, por lo tanto, estable si y sólo si $\rho(A) \leq 1$.

Es posible demostrar que los eigenvalores de A (consulte el ejercicio 15) son

$$\mu_i = 1 - 4\lambda\left(\operatorname{sen}\left(\frac{i\pi}{2m}\right)\right)^2, \quad \text{para cada } i = 1, 2, \ldots, m-1.$$

Por lo tanto, la condición para estabilidad, se reduce al determinar si

$$\rho(A) = \max_{1 \le i \le m-1} \left| 1 - 4\lambda \left(\operatorname{sen}\left(\frac{i\pi}{2m} \right) \right)^2 \right| \le 1,$$

y esto se simplifica para

$$0 \le \lambda \left(\operatorname{sen}\left(\frac{i\pi}{2m} \right) \right)^2 \le \frac{1}{2}, \quad \text{para cada } i = 1, 2, \ldots, m-1.$$

La estabilidad requiere que esta condición de desigualdad se mantenga cuando $h \to 0$, o, de forma equivalente, cuando $m \to \infty$. El hecho de que

$$\lim_{m \to \infty} \left[\operatorname{sen}\left(\frac{(m-1)\pi}{2m} \right) \right]^2 = 1$$

significa que se presentará estabilidad sólo si $0 \le \lambda \le \frac{1}{2}$.

Por definición, $\lambda = \alpha^2(k/h^2)$, por lo tanto, esta desigualdad requiere seleccionar h y k de tal forma que

$$\alpha^2 \frac{k}{h^2} \le \frac{1}{2}.$$

En el ejemplo 1, tenemos $\alpha^2 = 1$, por lo que esta condición se satisface cuando $h = 0.1$ y $k = 0.0005$. Pero cuando se incrementó k a 0.01 sin el aumento correspondiente en h, la razón fue

$$\frac{0.01}{(0.1)^2} = 1 > \frac{1}{2},$$

y los problemas de estabilidad se volvieron inmediatamente drásticos.

Consistente con la terminología del capítulo 5, llamamos al método de diferencias progresivas **condicionalmente estable**. El método converge en la solución de la ecuación (12.6) con la velocidad de convergencia $O(k + h^2)$, siempre y cuando

$$\alpha^2 \frac{k}{h^2} \le \frac{1}{2}$$

y las condiciones de continuidad requeridas para la solución se cumplan. (Para una prueba detallada de este hecho, observe [IK], p. 502–505.)

Método de diferencias regresivas

Para obtener un método que es **incondicionalmente estable**, consideramos uno de diferencias implícitas que resulta del uso del cociente de diferencias regresivas para $(\partial u/\partial t)(x_i, t_j)$ en la forma

$$\frac{\partial u}{\partial t}(x_i, t_j) = \frac{u(x_i, t_j) - u(x_i, t_{j-1})}{k} + \frac{k}{2} \frac{\partial^2 u}{\partial t^2}(x_i, \mu_j),$$

donde μ_j está en (t_{j-1}, t_j). Al sustituir esta ecuación, junto con la ecuación (12.8) para $\partial^2 u/\partial x^2$, en la ecuación diferencial parcial obtenemos

$$\frac{u(x_i, t_j) - u(x_i, t_{j-1})}{k} - \alpha^2 \frac{u(x_{i+1}, t_j) - 2u(x_i, t_j) + u(x_{i-1}, t_j)}{h^2}$$
$$= -\frac{k}{2} \frac{\partial^2 u}{\partial t^2}(x_i, \mu_j) - \alpha^2 \frac{h^2}{12} \frac{\partial^4 u}{\partial x^4}(\xi_i, t_j),$$

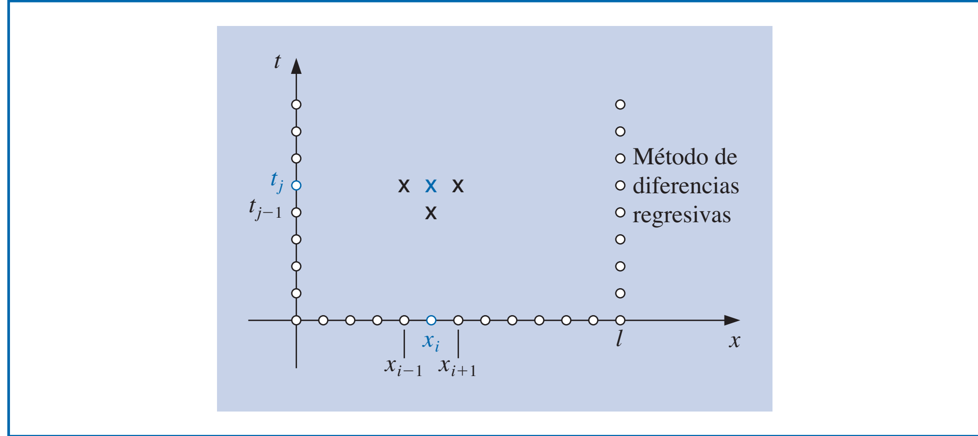
para algunas $\xi_i \in (x_{i-1}, x_{i+1})$. El **método de diferencias regresivas** resultante es

$$\frac{w_{ij} - w_{i,j-1}}{k} - \alpha^2 \frac{w_{i+1,j} - 2w_{ij} + w_{i-1,j}}{h^2} = 0, \tag{12.12}$$

para cada $i = 1, 2, \ldots, m-1$ y $j = 1, 2, \ldots$.

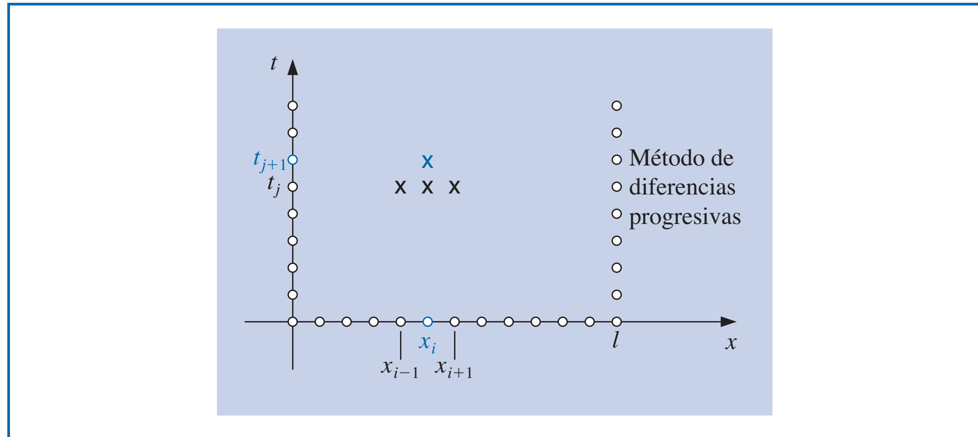
El método de diferencias regresivas implica los puntos de malla (x_i, t_{j-1}), (x_{i-1}, t_j), y (x_{i+1}, t_j) para aproximar el valor en (x_i, t_j), como se ilustra en la figura 12.9.

Figura 12.9



Puesto que las condiciones de frontera e iniciales relacionadas con el problema proporcionan información en los puntos de malla rodeados por un círculo, la figura no muestra procedimientos explícitos que se puedan utilizar para resolver la ecuación (12.12). Recuerde que en el método de diferencias progresivas (consulte la figura 12.10), las aproximaciones en (x_{i-1}, t_{j-1}), (x_i, t_{j-1}), y (x_{i+1}, t_{j-1}) se usaron para encontrar la aproximación en (x_i, t_j). Por lo que, se puede usar un método explícito para encontrar las aproximaciones con base en la información a partir de las condiciones iniciales y de frontera.

Figura 12.10



Si, de nuevo, permitimos que λ denote la cantidad $\alpha^2(k/h^2)$, el método de diferencia regresiva se vuelve

$$(1 + 2\lambda)w_{ij} - \lambda w_{i+1,j} - \lambda w_{i-1,j} = w_{i,j-1},$$

para cada $i = 1, 2, \ldots, m-1$ y $j = 1, 2, \ldots$. Por medio del conocimiento que $w_{i,0} = f(x_i)$, para cada $i = 1, 2, \ldots, m-1$ y $w_{m,j} = w_{0,j} = 0$, para cada $j = 1, 2, \ldots$, este método de diferencias tiene la representación matricial

$$\begin{bmatrix} (1+2\lambda) & -\lambda & 0 & \cdots & 0 \\ -\lambda & \ddots & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & \ddots & -\lambda \\ 0 & \cdots & 0 & -\lambda & (1+2\lambda) \end{bmatrix} \begin{bmatrix} w_{1,j} \\ w_{2,j} \\ \vdots \\ w_{m-1,j} \end{bmatrix} = \begin{bmatrix} w_{1,j-1} \\ w_{2,j-1} \\ \vdots \\ w_{m-1,j-1} \end{bmatrix}, \qquad (12.13)$$

o $A\mathbf{w}^{(j)} = \mathbf{w}^{(j-1)}$, para cada $i = 1, 2, \ldots$.

Por lo tanto, ahora debemos resolver el sistema lineal para obtener $\mathbf{w}^{(j)}$ a partir de $\mathbf{w}^{(j-1)}$. Sin embargo, $\lambda > 0$, por lo que la matriz A es definida positiva y estricta y diagonalmente dominante, así como tridiagonal. Por consiguiente, podemos usar tanto el algoritmo de factorización de Crout 6.7 como el algoritmo SOR 7.3 para resolver este sistema. El algoritmo 12.2 resuelve la ecuación (12.13) por medio de factorización de Crout, lo cual es aceptable a menos que m sea grande. En este algoritmo, suponemos, con el fin de paro, que se da una cota para t.

ALGORITMO **12.2**

Diferencias regresivas de la ecuación de calor

Para aproximar la solución de la ecuación diferencial parcial parabólica

$$\frac{\partial u}{\partial t}(x,t) - \alpha^2 \frac{\partial^2 u}{\partial x^2}(x,t) = 0, \quad 0 < x < l, \quad 0 < t < T,$$

sujeta a las condiciones de frontera

$$u(0,t) = u(l,t) = 0, \quad 0 < t < T,$$

y las condiciones iniciales

$$u(x,0) = f(x), \quad 0 \le x \le l:$$

ENTRADA extremo l; tiempo máximo T; constante α; enteros $m \ge 3$, $N \ge 1$.

SALIDA aproximaciones $w_{i,j}$ para $u(x_i, t_j)$ para cada $i = 1, \ldots, m-1$ y $j = 1, \ldots, N$.

Paso 1 Determine $h = l/m$;
$k = T/N$;
$\lambda = \alpha^2 k/h^2$.

Paso 2 Para $i = 1, \ldots, m-1$ determine $w_i = f(ih)$. (*Valores iniciales*.)
(*Los pasos* 3–11 *resuelven un sistema lineal tridiagonal por medio del algoritmo* 6.7.)

Paso 3 Determine $l_1 = 1 + 2\lambda$;
$u_1 = -\lambda/l_1$.

Paso 4 Para $i = 2, \ldots, m-2$ determine $l_i = 1 + 2\lambda + \lambda u_{i-1}$;
$u_i = -\lambda/l_i$.

Paso 5 Determine $l_{m-1} = 1 + 2\lambda + \lambda u_{m-2}$.

Paso 6 Para $j = 1, \ldots, N$ haga los pasos 7–11.

Paso 7 Determine $t = jk$; (*Actual* t_j.)
$z_1 = w_1/l_1$.

Paso 8 Para $i = 2, \ldots, m-1$ determine $z_i = (w_i + \lambda z_{i-1})/l_i$.

Paso 9 Determine $w_{m-1} = z_{m-1}$.

Paso 10 Para $i = m - 2, \ldots, 1$ determine $w_i = z_i - u_i w_{i+1}$.

Paso 11 SALIDA (t); (*Nota:* $t = t_j$.)
Para $i = 1, \ldots, m - 1$ determine $x = ih$;
SALIDA (x, w_i). (*Nota:* $w_i = w_{i,j}$.)

Paso 12 PARE. (*El procedimiento está completo*.) ■

Ejemplo 2 Use el método de diferencias regresivas (algoritmo 12.2) con $h = 0.1$ y $k = 0.01$ para aproximar la solución de la ecuación de calor

$$\frac{\partial u}{\partial t}(x,t) - \frac{\partial^2 u}{\partial x^2}(x,t) = 0, \quad 0 < x < 1, \quad 0 < t,$$

sujeta a las restricciones

$$u(0,t) = u(1,t) = 0, \quad 0 < t, \quad u(x,0) = \operatorname{sen} \pi x, \quad 0 \le x \le 1.$$

Solución Este problema se consideró en el ejemplo 1, donde encontramos que seleccionar $h = 0.1$ y $k = 0.0005$ da resultados bastante exactos. Sin embargo, con los valores en este ejemplo, $h = 0.1$ y $k = 0.01$, los resultados fueron excepcionalmente pobres. Para demostrar la estabilidad incondicional del método de diferencias regresivas, utilizaremos $h = 0.1$ y $k = 0.01$ y nuevamente, comparamos $w_{i,50}$ con $u(x_i, 0.5)$, donde $i = 0, 1, \ldots, 10$.

Los resultados mostrados en la tabla 12.4 tienen los mismos valores de h y k que en la quinta y sexta columnas de la tabla 12.3, lo cual ilustra la estabilidad de este método. ■

Tabla 12.4

x_i	$w_{i,50}$	$u(x_i, 0.5)$	$\lvert w_{i,50} - u(x_i, 0.5)\rvert$
0.0	0	0	
0.1	0.00289802	0.00222241	6.756×10^{-4}
0.2	0.00551236	0.00422728	1.285×10^{-3}
0.3	0.00758711	0.00581836	1.769×10^{-3}
0.4	0.00891918	0.00683989	2.079×10^{-3}
0.5	0.00937818	0.00719188	2.186×10^{-3}
0.6	0.00891918	0.00683989	2.079×10^{-3}
0.7	0.00758711	0.00581836	1.769×10^{-3}
0.8	0.00551236	0.00422728	1.285×10^{-3}
0.9	0.00289802	0.00222241	6.756×10^{-4}
1.0	0	0	

La razón por la que el método de diferencias regresivas no tiene los problemas de estabilidad del método de diferencias progresivas se puede observar al analizar los eigenvalores de la matriz A. Para el método de diferencias regresivas (consulte el ejercicio 16), los eigenvalores son

$$\mu_i = 1 + 4\lambda \left[\operatorname{sen}\left(\frac{i\pi}{2m}\right)\right]^2, \quad \text{para cada } i = 1, 2, \ldots, m-1.$$

Puesto que $\lambda > 0$ tenemos $\mu_i > 1$ para todas las $i = 1, 2, \ldots, m - 1$. Puesto que los eigenvalores de A^{-1} son los recíprocos de los de A, la relación espectral de A^{-1}, $\rho(A^{-1}) < 1$. Esto implica que A^{-1} es una matiz convergente.

Un error $\mathbf{e}^{(0)}$ en los datos iniciales produce un error $(A^{-1})^n \mathbf{e}^{(0)}$ en el enésimo paso del método de diferencias regresivas. Ya que A^{-1} es convergente,

$$\lim_{n \to \infty} (A^{-1})^n \mathbf{e}^{(0)} = \mathbf{0}.$$

Por lo que, el método es estable, independientemente de la selección de $\lambda = \alpha^2(k/h^2)$. En la terminología del capítulo 5, llamamos al método de diferencias regresivas un método **incondicionalmente estable**. El error de truncamiento local para el método es de orden

$O(k + h^2)$, siempre y cuando la solución de la ecuación diferencial satisfaga las condiciones de diferenciabilidad usual. En este caso, el método converge a la solución de la ecuación diferencial parcial con esta misma velocidad de convergencia (consulte [IK], p. 508).

La debilidad del método de diferencias regresivas resulta del hecho de que el error de truncamiento local tiene uno de orden $O(h^2)$ y otro de orden $O(k)$. Esto requiere hacer que los intervalos de tiempo sean mucho más pequeños que los intervalos del eje x. Sería claramente deseable tener un procedimiento con error de truncamiento local de orden $O(k^2 + h^2)$. El primer paso en esta dirección es usar una ecuación de diferencias que tiene error $O(k^2)$ para $u_t(x, t)$, en lugar de los que hemos utilizado previamente, cuyo error era $O(k)$. Esto se puede hacer usando la serie de Taylor en t para la función $u(x, t)$ en el punto (x_i, t_j) y al evaluar en (x_i, t_{j+1}) y (x_i, t_{j-1}) para obtener la fórmula de diferencias centradas

$$\frac{\partial u}{\partial t}(x_i, t_j) = \frac{u(x_i, t_{j+1}) - u(x_i, t_{j-1})}{2k} + \frac{k^2}{6}\frac{\partial^3 u}{\partial t^3}(x_i, \mu_j),$$

L. E. Richardson, quien, como observamos, está relacionado con la extrapolación, realizó un trabajo sustancial en la aproximación de ecuaciones diferenciales parciales.

donde $\mu_j \in (t_{j-1}, t_{j+1})$. El método de diferencias que resulta de sustituir esto y el cociente de diferencias usual para $(\partial^2 u/\partial x^2)$, la ecuación (12.8) en la ecuación diferencial recibe el nombre de **método de Richarson** y está dado por

$$\frac{w_{i,j+1} - w_{i,j-1}}{2k} - \alpha^2 \frac{w_{i+1,j} - 2w_{ij} + w_{i-1,j}}{h^2} = 0. \qquad (12.14)$$

Este método tiene error de truncamiento local de orden $O(k^2 + h^2)$, pero, por desgracia, al igual que el método de diferencia progresiva, tiene serios problemas de estabilidad (consulte los ejercicios 11 y 12).

Método de Crank-Nicolson

Un método más prometedor se deriva al promediar el método de diferencias progresivas en el j-ésimo paso en t,

$$\frac{w_{i,j+1} - w_{i,j}}{k} - \alpha^2 \frac{w_{i+1,j} - 2w_{i,j} + w_{i-1,j}}{h^2} = 0,$$

que tiene error de truncamiento local

$$\tau_F = \frac{k}{2}\frac{\partial^2 u}{\partial t^2}(x_i, \mu_j) + O(h^2),$$

y el método de diferencias regresivas en el $(j + 1)$ésimo paso en t,

$$\frac{w_{i,j+1} - w_{i,j}}{k} - \alpha^2 \frac{w_{i+1,j+1} - 2w_{i,j+1} + w_{i-1,j+1}}{h^2} = 0,$$

que tiene error de truncamiento local

$$\tau_B = -\frac{k}{2}\frac{\partial^2 u}{\partial t^2}(x_i, \hat{u}_j) + O(h^2).$$

Para continuar su trabajo como físico matemático durante la Segunda Guerra Mundial, John Crank (1916–2006) realizó investigaciones sobre la solución numérica de ecuaciones diferenciales parciales, en especial, problemas de conducción de calor. El método de Crank–Nicolson se basa en el trabajo realizado por Phyllis Nicolson (1917–1968), un físico en la Universidad Leeds. Su artículo original sobre el método apareció en 1947 [CN].

Si suponemos que

$$\frac{\partial^2 u}{\partial t^2}(x_i, \hat{\mu}_j) \approx \frac{\partial^2 u}{\partial t^2}(x_i, \mu_j),$$

entonces, el método de diferencia promediado,

$$\frac{w_{i,j+1} - w_{ij}}{k} - \frac{\alpha^2}{2}\left[\frac{w_{i+1,j} - 2w_{i,j} + w_{i-1,j}}{h^2} + \frac{w_{i+1,j+1} - 2w_{i,j+1} + w_{i-1,j+1}}{h^2}\right] = 0,$$

tiene error de truncamiento local de orden $O(k^2 + h^2)$, siempre y cuando, por supuesto, se satisfagan las condiciones de diferenciabilidad comunes.

Esto se conoce como el **método de Crank-Nicolson** y se representa en forma de matriz

$$A\mathbf{w}^{(j+1)} = B\mathbf{w}^{(j)}, \quad \text{para cada } j = 0, 1, 2, \ldots, \tag{12.15}$$

donde

$$\lambda = \alpha^2 \frac{k}{h^2}, \quad \mathbf{w}^{(j)} = (w_{1,j}, w_{2,j}, \ldots, w_{m-1,j})^t,$$

y las matrices A y B están dadas por:

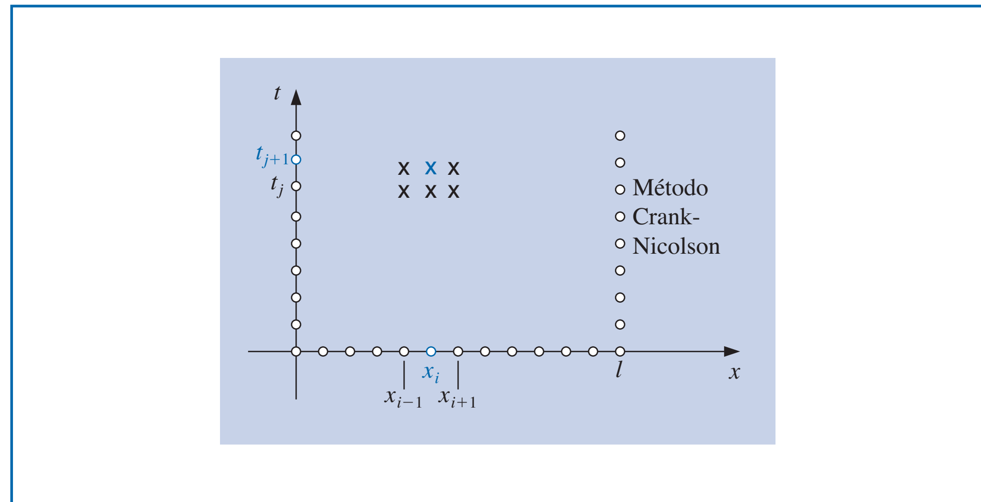
$$A = \begin{bmatrix} (1+\lambda) & -\frac{\lambda}{2} & 0 & \cdots & 0 \\ -\frac{\lambda}{2} & \ddots & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & -\frac{\lambda}{2} \\ 0 & \cdots & 0 & -\frac{\lambda}{2} & (1+\lambda) \end{bmatrix}$$

y

$$B = \begin{bmatrix} (1-\lambda) & \frac{\lambda}{2} & 0 & \cdots & 0 \\ \frac{\lambda}{2} & \ddots & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & \frac{\lambda}{2} \\ 0 & \cdots & 0 & \frac{\lambda}{2} & (1-\lambda) \end{bmatrix}.$$

La matriz no singular A es definida positiva, estricta y diagonalmente dominante y tridiagonal. Se puede usar tanto la factorización de Crout 6.7 como el algoritmo SOR 7.3 para obtener $\mathbf{w}^{(j+1)}$ a partir de $\mathbf{w}^{(j)}$, para cada $j = 0, 1, 2, \ldots$. El algoritmo 12.3 incorpora la factorización de Crout en la técnica Crank-Nicolson. Como en el algoritmo 12.2, una longitud finita para el intervalo de tiempo se debe especificar para determinar un procedimiento de detención. La verificación de que el método de Crank-Nicolson es incondicionalmente estable y tiene orden de convergencia $O(k^2 + h^2)$ se puede encontrar en [IK], p. 508–512. Un diagrama que muestra la interacción de los nodos para determinar una aproximación en (x_i, t_{j+1}) se muestra en la figura 12.11.

Figura 12.11



ALGORITMO
12.3

Método Crank-Nicolson

Para aproximar la solución de la ecuación diferencial parcial parabólica

$$\frac{\partial u}{\partial t}(x,t) - \alpha^2 \frac{\partial^2 u}{\partial x^2}(x,t) = 0, \quad 0 < x < l, \quad 0 < t < T,$$

sujeto a las condiciones de frontera

$$u(0,t) = u(l,t) = 0, \quad 0 < t < T,$$

y las condiciones iniciales

$$u(x,0) = f(x), \quad 0 \leq x \leq l:$$

ENTRADA extremo l; tiempo máximo T; constante α; enteros $m \geq 3$, $N \geq 1$.

SALIDA aproximaciones $w_{i,j}$ para $u(x_i, t_j)$ para cada $i = 1, \ldots, m-1$ y $j = 1, \ldots, N$..

Paso 1 Determine $h = l/m$;
$k = T/N$;
$\lambda = \alpha^2 k/h^2$;
$w_m = 0$.

Paso 2 Para $i = 1, \ldots, m-1$ determine $w_i = f(ih)$. (*Valores iniciales*.)
(*Los pasos* 3–11 *resuelven un sistema lineal tridiagonal por medio del algoritmo* 6.7.)

Paso 3 Determine $l_1 = 1 + \lambda$;
$u_1 = -\lambda/(2l_1)$.

Paso 4 Para $i = 2, \ldots, m-2$ determine $l_i = 1 + \lambda + \lambda u_{i-1}/2$;
$u_i = -\lambda/(2l_i)$.

Paso 5 Determine $l_{m-1} = 1 + \lambda + \lambda u_{m-2}/2$.

Paso 6 Para $j = 1, \ldots, N$ haga los pasos 7–11.

Paso 7 Determine $t = jk$; (*Actual* t_j.)

$$z_1 = \left[(1-\lambda)w_1 + \frac{\lambda}{2}w_2\right] \Big/ l_1.$$

Paso 8 Para $i = 2, \ldots, m-1$ determine

$$z_i = \left[(1-\lambda)w_i + \frac{\lambda}{2}(w_{i+1} + w_{i-1} + z_{i-1})\right] \Big/ l_i.$$

Paso 9 Determine $w_{m-1} = z_{m-1}$.

Paso 10 Para $i = m-2, \ldots, 1$ determine $w_i = z_i - u_i w_{i+1}$.

Paso 11 SALIDA (t); (*Nota:* $t = t_j$.)
Para $i = 1, \ldots, m-1$ establezca $x = ih$;
SALIDA (x, w_i). (*Nota:* $w_i = w_{i,j}$.)

Paso 12 PARE. (*El procedimiento está completo*.) ■

Ejemplo 3 Use el método Crank-Nicolson con $h = 0.1$ y $k = 0.01$ para aproximar la solución del problema

$$\frac{\partial u}{\partial t}(x, t) - \frac{\partial^2 u}{\partial x^2}(x, t) = 0, \quad 0 < x < 1 \quad 0 < t,$$

sujeto a las condiciones

$$u(0, t) = u(1, t) = 0, \quad 0 < t,$$

y

$$u(x, 0) = \text{sen}(\pi x), \quad 0 \le x \le 1.$$

Solución Al seleccionar $h = 0.1$ y $k = 0.01$ obtenemos $m = 10$, $N = 50$, y $\lambda = 1$) en el algoritmo 12.3. Recuerde que el método de diferencia progresiva provee resultados drásticamente pobres para esta selección de h y k, pero el método de diferencias regresivas da resultados que eran exactos alrededor de 2×10^{-3} para entradas en medio de la tabla. Los resultados en la tabla 12.5 indican el incremento de exactitud del método Crank-Nicolson sobre el método de diferencias regresivas, la mejor de las dos técnicas previamente analizadas. ■

Tabla 12.5

x_i	$w_{i,50}$	$u(x_i, 0.5)$	$\lvert w_{i,50} - u(x_i, 0.5)\rvert$
0.0	0	0	
0.1	0.00230512	0.00222241	8.271×10^{-5}
0.2	0.00438461	0.00422728	1.573×10^{-4}
0.3	0.00603489	0.00581836	2.165×10^{-4}
0.4	0.00709444	0.00683989	2.546×10^{-4}
0.5	0.00745954	0.00719188	2.677×10^{-4}
0.6	0.00709444	0.00683989	2.546×10^{-4}
0.7	0.00603489	0.00581836	2.165×10^{-4}
0.8	0.00438461	0.00422728	1.573×10^{-4}
0.9	0.00230512	0.00222241	8.271×10^{-5}
1.0	0	0	

La sección Conjunto de ejercicios 12.2 está disponible en línea. Encuentre la ruta de acceso en las páginas preliminares.

12.3 Ecuaciones diferenciales parciales hiperbólicas

En esta sección consideramos la solución numérica de la **ecuación de onda**, un ejemplo de una ecuación diferencial parcial *hiperbólica.* La ecuación de onda está dada por la ecuación diferencial

$$\frac{\partial^2 u}{\partial t^2}(x, t) - \alpha^2 \frac{\partial^2 u}{\partial x^2}(x, t) = 0, \quad 0 < x < l, \quad t > 0, \tag{12.16}$$

sujeta a las condiciones

$$u(0, t) = u(l, t) = 0, \quad \text{para} \quad t > 0,$$

$$u(x, 0) = f(x), \;\; \text{y} \;\; \frac{\partial u}{\partial t}(x, 0) = g(x), \quad \text{para} \quad 0 \le x \le l,$$

donde α es una constante dependiente de las condiciones físicas del problema.

Seleccione un entero $m > 0$ para definir los puntos de cuadrícula del eje x haciendo $h = l/m$. Además, seleccione un tamaño de longitud de paso-tiempo $k > 0$. Los puntos de malla (x_i, t_j) están definidos por

$$x_i = ih \quad \text{y} \quad t_j = jk,$$

para cada $i = 0, 1, \ldots, m$ y $j = 0, 1, \ldots$.

En cualquier punto de malla interior (x_i, t_j), la ecuación de onda se vuelve

$$\frac{\partial^2 u}{\partial t^2}(x_i, t_j) - \alpha^2 \frac{\partial^2 u}{\partial x^2}(x_i, t_j) = 0. \tag{12.17}$$

El método de diferencias se obtiene a través del cociente de diferencias centradas para las segundas derivadas parciales dadas por

$$\frac{\partial^2 u}{\partial t^2}(x_i, t_j) = \frac{u(x_i, t_{j+1}) - 2u(x_i, t_j) + u(x_i, t_{j-1})}{k^2} - \frac{k^2}{12}\frac{\partial^4 u}{\partial t^4}(x_i, \mu_j),$$

donde $\mu_j \in (t_{j-1}, t_{j+1})$, y

$$\frac{\partial^2 u}{\partial x^2}(x_i, t_j) = \frac{u(x_{i+1}, t_j) - 2u(x_i, t_j) + u(x_{i-1}, t_j)}{h^2} - \frac{h^2}{12}\frac{\partial^4 u}{\partial x^4}(\xi_i, t_j),$$

donde $\xi_i \in (x_{i-1}, x_{i+1})$. Al sustituir éstas en la ecuación (12.17) obtenemos

$$\frac{u(x_i, t_{j+1}) - 2u(x_i, t_j) + u(x_i, t_{j-1})}{k^2} - \alpha^2 \frac{u(x_{i+1}, t_j) - 2u(x_i, t_j) + u(x_{i-1}, t_j)}{h^2}$$
$$= \frac{1}{12}\left[k^2 \frac{\partial^4 u}{\partial t^4}(x_i, \mu_j) - \alpha^2 h^2 \frac{\partial^4 u}{\partial x^4}(\xi_i, t_j)\right].$$

Al ignorar el término de error

$$\tau_{i,j} = \frac{1}{12}\left[k^2 \frac{\partial^4 u}{\partial t^4}(x_i, \mu_j) - \alpha^2 h^2 \frac{\partial^4 u}{\partial x^4}(\xi_i, t_j)\right], \tag{12.18}$$

obtenemos la ecuación de diferencias

$$\frac{w_{i,j+1} - 2w_{i,j} + w_{i,j-1}}{k^2} - \alpha^2 \frac{w_{i+1,j} - 2w_{i,j} + w_{i-1,j}}{h^2} = 0.$$

Defina $\lambda = \alpha k/h$. Entonces, podemos escribir la ecuación de diferencia como

$$w_{i,j+1} - 2w_{i,j} + w_{i,j-1} - \lambda^2 w_{i+1,j} + 2\lambda^2 w_{i,j} - \lambda^2 w_{i-1,j} = 0$$

y resolver para $w_{i,j+1}$, la aproximación de longitud de paso-tiempo, para obtener

$$w_{i,j+1} = 2(1 - \lambda^2)w_{i,j} + \lambda^2(w_{i+1,j} + w_{i-1,j}) - w_{i,j-1}. \tag{12.19}$$

Esta ecuación se mantiene para cada $i = 1, 2, \ldots, m-1$ y $j = 1, 2, \ldots$. Las condiciones de frontera proporcionan

$$w_{0,j} = w_{m,j} = 0, \quad \text{para cada } j = 1, 2, 3, \ldots, \tag{12.20}$$

y la condición inicial implica que

$$w_{i,0} = f(x_i), \quad \text{para cada } i = 1, 2, \ldots, m - 1. \tag{12.21}$$

Escribir este conjunto de ecuaciones en forma matricial nos da

$$\begin{bmatrix} w_{1,j+1} \\ w_{2,j+1} \\ \vdots \\ w_{m-1,j+1} \end{bmatrix} = \begin{bmatrix} 2(1-\lambda^2) & \lambda^2 & 0 & \cdots & 0 \\ \lambda^2 & 2(1-\lambda^2) & \lambda^2 & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & \ddots & \lambda^2 \\ 0 & \cdots & 0 & \lambda^2 & 2(1-\lambda^2) \end{bmatrix} \begin{bmatrix} w_{1,j} \\ w_{2,j} \\ \vdots \\ w_{m-1,j} \end{bmatrix} - \begin{bmatrix} w_{1,j-1} \\ w_{2,j-1} \\ \vdots \\ w_{m-1,j-1} \end{bmatrix}. \tag{12.22}$$

Las ecuaciones (12.19) y (12.22) implican que el $(j+1)$-ésimo paso requiere valores desde el j-ésimo y $(j-1)$-ésimo pasos. (Consulte la figura 12.12.) Esto produce un pequeño problema inicial porque los valores de $j = 0$ están dados por la ecuación (12.21), pero los valores para $j = 1$, necesarios en la ecuación (12.19) para calcular $w_{i,2}$, deben obtenerse a partir de la condición de velocidad inicial.

$$\frac{\partial u}{\partial t}(x, 0) = g(x), \quad 0 \le x \le l.$$

Figura 12.12

Un enfoque es reemplazar $\partial u/\partial t$ mediante una aproximación de diferencias progresiva

$$\frac{\partial u}{\partial t}(x_i, 0) = \frac{u(x_i, t_1) - u(x_i, 0)}{k} - \frac{k}{2}\frac{\partial^2 u}{\partial t^2}(x_i, \tilde{\mu}_i), \tag{12.23}$$

para algún $\tilde{\mu}_i$ en $(0, t_1)$. Al resolver para $u(x_i, t_1)$ en la ecuación obtenemos

$$\begin{aligned} u(x_i, t_1) &= u(x_i, 0) + k\frac{\partial u}{\partial t}(x_i, 0) + \frac{k^2}{2}\frac{\partial^2 u}{\partial t^2}(x_i, \tilde{\mu}_i) \\ &= u(x_i, 0) + kg(x_i) + \frac{k^2}{2}\frac{\partial^2 u}{\partial t^2}(x_i, \tilde{\mu}_i). \end{aligned}$$

Al borrar el término de truncamiento, obtenemos la aproximación,

$$w_{i,1} = w_{i,0} + kg(x_i), \quad \text{para cada } i = 1, \ldots, m-1. \tag{12.24}$$

Sin embargo, esta aproximación tiene error de truncamiento de sólo $O(k)$, mientras el error de truncamiento en la ecuación (12.18) es $O(k^2)$.

Mejora de la aproximación inicial

Para obtener una mejor aproximación para $u(x_i, 0)$, expanda $u(x_i, t_1)$ en un segundo polinomio Maclaurin en t. Entonces

$$u(x_i, t_1) = u(x_i, 0) + k\frac{\partial u}{\partial t}(x_i, 0) + \frac{k^2}{2}\frac{\partial^2 u}{\partial t^2}(x_i, 0) + \frac{k^3}{6}\frac{\partial^3 u}{\partial t^3}(x_i, \hat{\mu}_i),$$

para algunas $\hat{\mu}_i$ en $(0, t_1)$. Si f'' existe, entonces

$$\frac{\partial^2 u}{\partial t^2}(x_i, 0) = \alpha^2\frac{\partial^2 u}{\partial x^2}(x_i, 0) = \alpha^2\frac{d^2 f}{dx^2}(x_i) = \alpha^2 f''(x_i)$$

y

$$u(x_i, t_1) = u(x_i, 0) + kg(x_i) + \frac{\alpha^2 k^2}{2}f''(x_i) + \frac{k^3}{6}\frac{\partial^3 u}{\partial t^3}(x_i, \hat{\mu}_i).$$

Esto produce una aproximación con error $O(k^3)$:

$$w_{i1} = w_{i0} + kg(x_i) + \frac{\alpha^2 k^2}{2}f''(x_i).$$

Si $f \in C^4[0, 1]$ pero $f''(x_i)$ no está disponible fácilmente, podemos usar la ecuación de diferencias en la ecuación (4.9) para escribir

$$f''(x_i) = \frac{f(x_{i+1}) - 2f(x_i) + f(x_{i-1})}{h^2} - \frac{h^2}{12}f^{(4)}(\tilde{\xi}_i),$$

para algunas $\tilde{\xi}_i$ en (x_{i-1}, x_{i+1}). Esto implica que

$$u(x_i, t_1) = u(x_i, 0) + kg(x_i) + \frac{k^2\alpha^2}{2h^2}[f(x_{i+1}) - 2f(x_i) + f(x_{i-1})] + O(k^3 + h^2k^2).$$

Puesto que $\lambda = k\alpha/h$, podemos escribir esto como

$$\begin{aligned}u(x_i, t_1) &= u(x_i, 0) + kg(x_i) + \frac{\lambda^2}{2}[f(x_{i+1}) - 2f(x_i) + f(x_{i-1})] + O(k^3 + h^2k^2)\\ &= (1-\lambda^2)f(x_i) + \frac{\lambda^2}{2}f(x_{i+1}) + \frac{\lambda^2}{2}f(x_{i-1}) + kg(x_i) + O(k^3 + h^2k^2).\end{aligned}$$

Por lo tanto, la ecuación de diferencias,

$$w_{i,1} = (1-\lambda^2)f(x_i) + \frac{\lambda^2}{2}f(x_{i+1}) + \frac{\lambda^2}{2}f(x_{i-1}) + kg(x_i), \quad (12.25)$$

se puede $w_{i,1}$ usar para encontrar $i = 1, 2, \ldots, m-1$. Para determinar aproximaciones subsiguientes, usamos el sistema en la ecuación (12.22).

El algoritmo 12.4 usa la ecuación (12.25) para aproximar $w_{i,1}$, a pesar de que la ecuación (12.24) también podría usarse. Se supone que existe una cota superior para el valor de t que se puede utilizar en la técnica de detención y que $k = T/N$, donde N también está dada.

ALGORITMO
12.4

Diferencias finitas para la ecuación de onda

Para aproximar la solución de la ecuación de onda

$$\frac{\partial^2 u}{\partial t^2}(x,t) - \alpha^2 \frac{\partial^2 u}{\partial x^2}(x,t) = 0, \quad 0 < x < l, \quad 0 < t < T,$$

sujeta las condiciones de frontera

$$u(0,t) = u(l,t) = 0, \quad 0 < t < T,$$

y las condiciones iniciales

$$u(x,0) = f(x), \quad \text{y} \quad \frac{\partial u}{\partial t}(x,0) = g(x), \quad \text{para} \quad 0 \le x \le l,$$

ENTRADA extremo l; tiempo máximo T; constante α; enteros $m \ge 2$, $N \ge 2$.

SALIDA aproximaciones $w_{i,j}$ para $u(x_i, t_j)$ para cada $i = 0, \ldots, m$ y $j = 0, \ldots, N$.

Paso 1 Determine $h = l/m$;
$k = T/N$;
$\lambda = k\alpha/h$.

Paso 2 Para $j = 1, \ldots, N$ determine $w_{0,j} = 0$;
$w_{m,j} = 0$;

Paso 3 Determine $w_{0,0} = f(0)$;
$w_{m,0} = f(l)$.

Paso 4 Para $i = 1, \ldots, m-1$ (*Inicialice para* $t = 0$ *y* $t = k$.)
determine $w_{i,0} = f(ih)$;
$w_{i,1} = (1-\lambda^2)f(ih) + \frac{\lambda^2}{2}[f((i+1)h) + f((i-1)h)] + kg(ih)$.

Paso 5 Para $j = 1, \ldots, N-1$ (*Realice la multiplicación de matriz.*)
para $i = 1, \ldots, m-1$
determine $w_{i,j+1} = 2(1-\lambda^2)w_{i,j} + \lambda^2(w_{i+1,j} + w_{i-1,j}) - w_{i,j-1}$.

Paso 6 Para $j = 0, \ldots, N$
determine $t = jk$;
para $i = 0, \ldots, m$
determine $x = ih$;
SALIDA $(x, t, w_{i,j})$.

Paso 7 PARE. (*El procedimiento está completo.*) ■

Ejemplo 1 Aproxime la solución del problema hiperbólico

$$\frac{\partial^2 u}{\partial t^2}(x,t) - 4\frac{\partial^2 u}{\partial x^2}(x,t) = 0, \quad 0 < x < 1, \quad 0 < t,$$

con condiciones de frontera

$$u(0,t) = u(1,t) = 0, \quad \text{para } 0 < t,$$

y condiciones iniciales

$$u(x,0) = \operatorname{sen}(\pi x), \quad 0 \le x \le 1, \quad \text{y} \quad \frac{\partial u}{\partial t}(x,0) = 0, \quad 0 \le x \le 1,$$

Usando $h = 0.1$ y $k = 0.05$. Compare los resultados con la solución exacta

$$u(x, t) = \operatorname{sen} \pi x \cos 2\pi t.$$

Solución Al seleccionar $h = 0.1$ y $k = 0.05$ da $\lambda = 1, m = 10$, y $N = 20$. Seleccionaremos un tiempo máximo $T = 1$ y aplicaremos el algoritmo de diferencias finitas 12.4. Esto produce las aproximaciones $w_{i,N}$ a $u(0.1i, 1)$ para $i = 0, 1, \ldots, 10$. Estos resultados se muestran en la tabla 12.6 y son correctos para los lugares dados. ■

Tabla 12.6

x_i	$w_{i,20}$
0.0	0.0000000000
0.1	0.3090169944
0.2	0.5877852523
0.3	0.8090169944
0.4	0.9510565163
0.5	1.0000000000
0.6	0.9510565163
0.7	0.8090169944
0.8	0.5877852523
0.9	0.3090169944
1.0	0.0000000000

Los resultados del ejemplo eran muy exactos, más de lo que el error de truncamiento $O(k^2 + h^2)$ nos permitiría creer. Esto es porque la verdadera solución para la ecuación es infinitamente diferenciable. Cuando éste es el caso, la serie de Taylor da

$$\frac{u(x_{i+1}, t_j) - 2u(x_i, t_j) + u(x_{i-1}, t_j)}{h^2} = \frac{\partial^2 u}{\partial x^2}(x_i, t_j) + 2\left[\frac{h^2}{4!}\frac{\partial^4 u}{\partial x^4}(x_i, t_j) + \frac{h^4}{6!}\frac{\partial^6 u}{\partial x^6}(x_i, t_j) + \cdots\right]$$

y

$$\frac{u(x_i, t_{j+1}) - 2u(x_i, t_j) + u(x_i, t_{j-1})}{k^2} = \frac{\partial^2 u}{\partial t^2}(x_i, t_j) + 2\left[\frac{k^2}{4!}\frac{\partial^4 u}{\partial t^4}(x_i, t_j) + \frac{h^4}{6!}\frac{\partial^6 u}{\partial t^6}(x_i, t_j) + \cdots\right].$$

Puesto que $u(x, t)$ satisface la ecuación diferencial parcial,

$$\frac{u(x_i, t_{j+1}) - 2u(x_i, t_j) + u(x_i, t_{j-1})}{k^2} - \alpha^2 \frac{u(x_{i+1}, t_j) - 2u(x_i, t_j) + u(x_{i-1}, t_j)}{h^2}$$
$$= 2\left[\frac{1}{4!}\left(k^2 \frac{\partial^4 u}{\partial t^4}(x_i, t_j) - \alpha^2 h^2 \frac{\partial^4 u}{\partial x^4}(x_i, t_j)\right) + \frac{1}{6!}\left(k^4 \frac{\partial^6 u}{\partial t^6}(x_i, t_j) - \alpha^2 h^4 \frac{\partial^6 u}{\partial x^6}(x_i, t_j)\right) + \cdots\right]. \quad (12.26)$$

Sin embargo, al diferenciar la ecuación de onda obtenemos

$$k^2 \frac{\partial^4 u}{\partial t^4}(x_i, t_j) = k^2 \frac{\partial^2}{\partial t^2}\left[\alpha^2 \frac{\partial^2 u}{\partial x^2}(x_i, t_j)\right] = \alpha^2 k^2 \frac{\partial^2}{\partial x^2}\left[\frac{\partial^2 u}{\partial t^2}(x_i, t_j)\right]$$
$$= \alpha^2 k^2 \frac{\partial^2}{\partial x^2}\left[\alpha^2 \frac{\partial^2 u}{\partial x^2}(x_i, t_j)\right] = \alpha^4 k^2 \frac{\partial^4 u}{\partial x^4}(x_i, t_j),$$

y observamos que puesto que $\lambda^2 = (\alpha^2 k^2/h^2) = 1$, tenemos

$$\frac{1}{4!}\left[k^2 \frac{\partial^4 u}{\partial t^4}(x_i, t_j) - \alpha^2 h^2 \frac{\partial^4 u}{\partial x^4}(x_i, t_j)\right] = \frac{\alpha^2}{4!}[\alpha^2 k^2 - h^2]\frac{\partial^4 u}{\partial x^4}(x_i, t_j) = 0.$$

Al continuar de esta forma, todos los términos en el lado derecho de la ecuación (12.26) son 0, lo que implica que el error de truncamiento local es 0. Los únicos errores en el ejemplo 1 son los que se deben a la aproximación de $w_{i,1}$ y al de redondeo.

Como en el caso del método de diferencias progresivas para la ecuación de calor el método de diferencias finitas explícita para la ecuación de onda tiene problemas de estabilidad. De hecho, es necesario que $\lambda = \alpha k/h \leq 1$ para que el método sea estable. (Consulte [IK], p. 489.) El método explícito dado en el algoritmo 12.4, con $\lambda \leq 1$ es $O(h^2 + k^2)$ convergente si f y g son suficientemente diferenciables. Para verificarlo consulte [IK], p. 491.

A pesar de que no los analizaremos, existen métodos implícitos que son incondicionalmente estables. Un análisis de estos métodos se puede encontrar en [Am], p. 199, [Mi] o [Sm, B].

La sección Conjunto de ejercicios 12.3 está disponible en línea. Encuentre la ruta de acceso en las páginas preliminares.



12.4 Una introducción al método de elementos finitos

Los elementos finitos comenzaron en la década de 1950 en la industria de las aeronaves. Al uso de las técnicas le siguió un artículo de Turner, Clough, Martin, y Topp [TCMT] que fue publicado en 1956. La amplia aplicación de los métodos requería grandes recursos computacionales que no estuvieron disponibles hasta principios de la década de 1970.

El **método de elementos finitos** es similar al método de Rayleigh-Ritz para aproximar la solución de los problemas de valor en la frontera para dos puntos que se presentó en la sección 11.5. Fue desarrollado originalmente para su uso en ingeniería civil, pero ahora son útiles en la aproximación de las soluciones para las ecuaciones diferenciales parciales que surgen en todas las áreas de las matemáticas aplicadas.

Una ventaja que tiene el método de elementos finitos sobre los métodos de diferencias finitas es la relativa facilidad con la que se manejan las condiciones de frontera del problema. Muchos problemas físicos tienen condiciones de frontera que implican derivadas y fronteras de formas irregulares. Dichas condiciones son difíciles de manejar con técnicas de diferencias finitas porque cada condición en la frontera que implica una derivada se debe aproximar mediante un cociente de diferencias en los puntos de cuadrícula y una forma irregular de la frontera hace que colocar los puntos de la cuadrícula sea difícil. El método de elementos finitos incluye las condiciones de frontera como integrales en una función que se está minimizando, por lo que el procedimiento de construcción es independiente de las condiciones de frontera particulares del problema.

En nuestro análisis, consideramos la ecuación diferencial parcial

$$\frac{\partial}{\partial x}\left(p(x, y)\frac{\partial u}{\partial x}\right) + \frac{\partial}{\partial y}\left(q(x, y)\frac{\partial u}{\partial y}\right) + r(x, y)u(x, y) = f(x, y), \qquad (12.27)$$

con $(x, y) \in \mathcal{D}$, donde $\mathcal{D}$ es una región plana con frontera $\mathcal{S}$.

Las condiciones de frontera de la forma

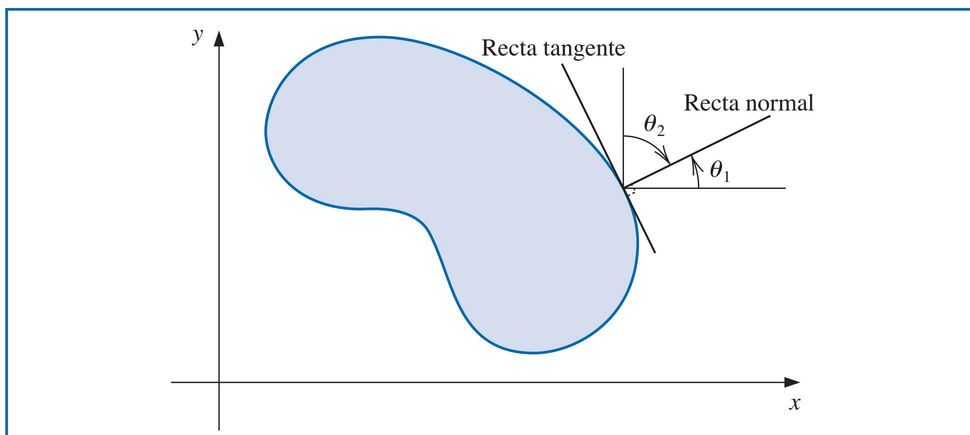
$$u(x, y) = g(x, y) \qquad (12.28)$$

se imponen en una parte, $\mathcal{S}_1$, de la frontera. En el resto la frontera, $\mathcal{S}_2$, se requiere la solución $u(x, y)$ para satisfacer

$$p(x, y)\frac{\partial u}{\partial x}(x, y)\cos\theta_1 + q(x, y)\frac{\partial u}{\partial y}(x, y)\cos\theta_2 + g_1(x, y)u(x, y) = g_2(x, y), \qquad (12.29)$$

donde θ_1 y θ_2 son los ángulos de dirección de la normal exterior para la frontera en el punto (x, y). (Consulte la figura 12.13).

Figura 12.13



Los problemas físicos en las áreas de la mecánica sólida y la elasticidad tienen ecuaciones diferenciales parciales relacionadas similares a la ecuación (12.27). En general, la solución del problema de este tipo minimiza cierta función, que implica integrales, sobre una clase de funciones determinadas por el problema.

Suponga que p, q, r y f son todas continuas en $\mathcal{D} \cup \mathcal{S}$, y que q tiene primeras derivadas parciales continuas y g_1 y g_2 son continuas en $\mathcal{S}_2$. Suponga, además, que $p(x, y)$ $0, q(x, y) > 0, r(x, y) \leq 0$, y $g_1(x, y) > 0$. Entonces, una solución de la ecuación (12.27) solamente minimiza la función

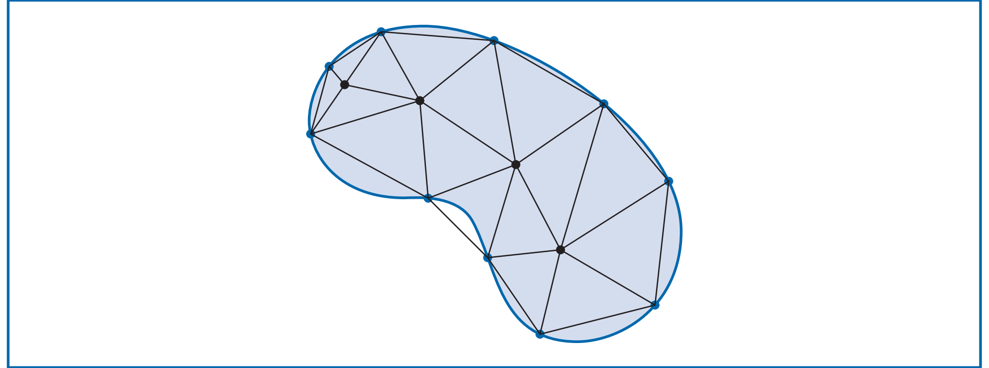
$$I[w] = \iint_{\mathcal{D}} \left\{ \frac{1}{2} \left[p(x, y) \left(\frac{\partial w}{\partial x} \right)^2 + q(x, y) \left(\frac{\partial w}{\partial y} \right)^2 - r(x, y) w^2 \right] + f(x, y) w \right\} dx\, dy$$
$$+ \int_{\mathcal{S}_2} \left\{ -g_2(x, y) w + \frac{1}{2} g_1(x, y) w^2 \right\} dS \tag{12.30}$$

sobre todas las funciones dos veces continuamente diferenciables w que satisfacen la ecuación (12.28) en $\mathcal{S}_1$. El método de elementos finitos aproxima esta solución al minimizar la función I sobre una clase más pequeña de funciones, justo como el método de Rayleigh-Ritz para el problema de valor en la frontera considerado en la sección 11.5.

Definición de elementos

El primer paso es dividir la región en un número finito de secciones, o elementos, de una forma regular, ya sea rectángulos o triángulos. (Consulte la figura 12.14.)

Figura 12.14



En general, el conjunto de funciones que se usa para la aproximación es un conjunto de polinomios por tramos de grado fijo en x y y, y la aproximación requiere que los polinomios se reconstruyan de tal forma que la función resultante sea continua con una primera o segunda derivada integrable o continua en toda la región. Los polinomios de tipo lineal en x y y,

$$\phi(x, y) = a + bx + cy,$$

se usan comúnmente con elementos triangulares, mientras que los polinomios de tipo bilineal en x y y,

$$\phi(x, y) = a + bx + cy + dxy,$$

se utilizan con elementos rectangulares.

Suponga que la región $\mathcal{D}$ se ha subdividido en elementos triangulares. El conjunto de triángulos se denota como D y los vértices de estos triángulos reciben el nombre de **nodos**. El método busca una aproximación de la forma

$$\phi(x, y) = \sum_{i=1}^{m} \gamma_i \phi_i(x, y), \tag{12.31}$$

donde $\phi_1, \phi_2, \ldots, \phi_m$ son polinomios lineales por tramos linealmente independientes y $\gamma_1, \gamma_2, \ldots, \gamma_m$ son constantes. Algunas de estas constantes, por ejemplo, $\gamma_{n+1}, \gamma_{n+2}, \ldots, \gamma_m$, se usan para garantizar que la condición de frontera,

$$\phi(x, y) = g(x, y),$$

se satisfaga en $\mathcal{S}_1$ y las constantes restantes $\gamma_1, \gamma_2, \ldots, \gamma_n$, se usan para minimizar la función $I\left[\sum_{i=1}^{m} \gamma_i \phi_i\right]$.

Reemplazando la forma de $\phi(x, y)$ dada en la ecuación (12.31) para w en la ecuación (12.30) produce

$$\begin{aligned}
I[\phi] &= I\left[\sum_{i=1}^{m} \gamma_i \phi_i\right] \\
&= \iint_{\mathcal{D}} \left(\frac{1}{2}\left\{ p(x, y)\left[\sum_{i=1}^{m} \gamma_i \frac{\partial \phi_i}{\partial x}(x, y)\right]^2 + q(x, y)\left[\sum_{i=1}^{m} \gamma_i \frac{\partial \phi_i}{\partial y}(x, y)\right]^2 \right.\right. \\
&\quad \left.\left. - r(x, y)\left[\sum_{i=1}^{m} \gamma_i \phi_i(x, y)\right]^2 \right\} + f(x, y)\sum_{i=1}^{m} \gamma_i \phi_i(x, y) \right) dy\, dx \\
&\quad + \int_{\mathcal{S}_2} \left\{ -g_2(x, y)\sum_{i=1}^{m} \gamma_i \phi_i(x, y) + \frac{1}{2} g_1(x, y)\left[\sum_{i=1}^{m} \gamma_i \phi_i(x, y)\right]^2 \right\} dS. \quad (12.32)
\end{aligned}$$

Considere I como una función de $\gamma_1, \gamma_2, \ldots, \gamma_n$. Para que se presente un mínimo, debemos tener

$$\frac{\partial I}{\partial \gamma_j} = 0, \quad \text{para cada } j = 1, 2, \ldots, n.$$

Derivando (12.32) obtenemos

$$\begin{aligned}
\frac{\partial I}{\partial \gamma_j} &= \iint_{\mathcal{D}} \left\{ p(x, y)\sum_{i=1}^{m} \gamma_i \frac{\partial \phi_i}{\partial x}(x, y)\frac{\partial \phi_j}{\partial x}(x, y) \right. \\
&\quad + q(x, y)\sum_{i=1}^{m} \gamma_i \frac{\partial \phi_i}{\partial y}(x, y)\frac{\partial \phi_j}{\partial y}(x, y) \\
&\quad \left. - r(x, y)\sum_{i=1}^{m} \gamma_i \phi_i(x, y)\phi_j(x, y) + f(x, y)\phi_j(x, y) \right\} dx\, dy \\
&\quad + \int_{\mathcal{S}_2} \left\{ -g_2(x, y)\phi_j(x, y) + g_1(x, y)\sum_{i=1}^{m} \gamma_i \phi_i(x, y)\phi_j(x, y) \right\} dS,
\end{aligned}$$

por lo que

$$\begin{aligned}
0 = \sum_{i=1}^{m} &\left[\iint_{\mathcal{D}} \left\{ p(x, y)\frac{\partial \phi_i}{\partial x}(x, y)\frac{\partial \phi_j}{\partial x}(x, y) + q(x, y)\frac{\partial \phi_i}{\partial y}(x, y)\frac{\partial \phi_j}{\partial y}(x, y) \right.\right. \\
&\quad \left. - r(x, y)\phi_i(x, y)\phi_j(x, y) \right\} dx\, dy \\
&\quad \left. + \int_{\mathcal{S}_2} g_1(x, y)\phi_i(x, y)\phi_j(x, y)\, dS \right] \gamma_i \\
&\quad + \iint_{\mathcal{D}} f(x, y)\phi_j(x, y)\, dx\, dy - \int_{\mathcal{S}_2} g_2(x, y)\phi_j(x, y)\, dS,
\end{aligned}$$

para cada $j = 1, 2, \ldots, n$. Este conjunto de ecuaciones se puede escribir como un sistema lineal:

$$A\mathbf{c} = \mathbf{b},$$

donde $\mathbf{c} = (\gamma_1, \ldots, \gamma_n)^t$ y donde las matrices $n \times n$ $A = (\alpha_{ij})$ y $\mathbf{b} = (\beta_1, \ldots, \beta_n)^t$ están definidas por

$$\alpha_{ij} = \iint_{\mathcal{D}} \left[p(x, y)\frac{\partial \phi_i}{\partial x}(x, y)\frac{\partial \phi_j}{\partial x}(x, y) + q(x, y)\frac{\partial \phi_i}{\partial y}(x, y)\frac{\partial \phi_j}{\partial y}(x, y) \right. \tag{12.33}$$
$$\left. - r(x, y)\phi_i(x, y)\phi_j(x, y) \right] dx\, dy + \int_{\mathcal{S}_2} g_1(x, y)\phi_i(x, y)\phi_j(x, y)\, dS,$$

para cada $i = 1, 2, \ldots, n$ y $j = 1, 2, \ldots, m$,

$$\beta_i = -\iint_{\mathcal{D}} f(x, y)\phi_i(x, y)\, dx\, dy + \int_{\mathcal{S}_2} g_2(x, y)\phi_i(x, y)\, dS - \sum_{k=n+1}^{m} \alpha_{ik}\gamma_k, \tag{12.34}$$

para cada $i = 1, \ldots, n$.

La selección particular de las funciones base es importante porque, a menudo, la selección adecuada puede crear la matriz definida positiva y de banda A. Para el problema de segundo orden (12.27), suponemos que $\mathcal{D}$ es poligonal, de tal forma que $\mathcal{D} = D$, y que $\mathcal{S}$ es un conjunto contiguo de líneas rectas.

Triangulación de la región

Para comenzar el procedimiento, dividimos la región D en un conjunto de triángulos $T_1, T_2, \ldots, T_M$, con el i-ésimo triángulo con tres vértices, o nodos, denotados

$$V_j^{(i)} = \left(x_j^{(i)}, y_j^{(i)}\right), \quad \text{para } j = 1, 2, 3.$$

Para simplificar la notación, escribimos $V_j^{(i)}$ simplemente como $V_j = (x_j, y_j)$ cuando trabajamos con el triángulo fijo T_i. Con cada vértice V_j, asociamos el polinomio lineal

$$N_j^{(i)}(x, y) \equiv N_j(x, y) = a_j + b_j x + c_j y, \quad \text{donde} \quad N_j^{(i)}(x_k, y_k) = \begin{cases} 1, & \text{si } j = k, \\ 0, & \text{si } j \neq k. \end{cases}$$

Esto produce sistemas lineales de la forma

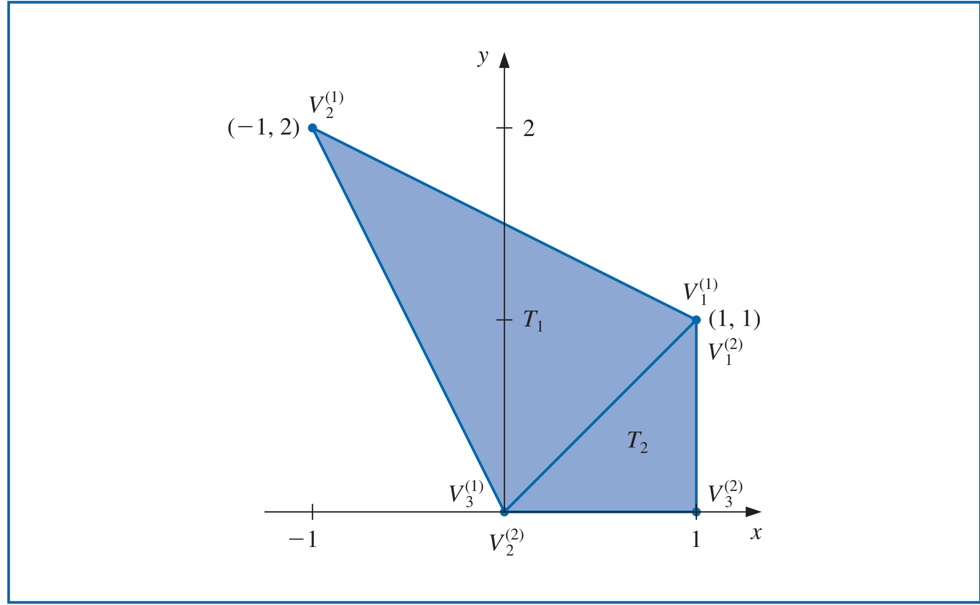
$$\begin{bmatrix} 1 & x_1 & y_1 \\ 1 & x_2 & y_2 \\ 1 & x_3 & y_3 \end{bmatrix} \begin{bmatrix} a_j \\ b_j \\ c_j \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix},$$

con el elemento 1 presente en la j-ésima fila en el vector de la derecha (aquí $j = 2$).

Sean $E_1, \ldots, E_n$ etiquetas de los nodos en $D \cup \mathcal{S}$. Con cada nodo E_k, asociamos una función ϕ_k que es lineal en cada triángulo, tiene el valor 1 en E_k y es 0 en cada uno de los otros nodos. Esta selección hace que ϕ_k sea idéntico a $N_j^{(i)}$ en el triángulo T_i cuando el nodo E_k es el vértice denotado $V_j^{(i)}$.

Ilustración Suponga que un problema de elementos finitos contiene los triángulos T_1 y T_2 mostrados en la figura 12.15.

Figura 12.15



La función lineal $N_1^{(1)}(x, y)$ asume el valor 1 en $(1, 1)$ y el valor 0 tanto en $(0, 0)$ como en $(-1, 2)$ satisface

$$a_1^{(1)} + b_1^{(1)}(1) + c_1^{(1)}(1) = 1,$$
$$a_1^{(1)} + b_1^{(1)}(-1) + c_1^{(1)}(2) = 0,$$

y

$$a_1^{(1)} + b_1^{(1)}(0) + c_1^{(1)}(0) = 0.$$

La solución a este sistema es $a_1^{(1)} = 0$, $b_1^{(1)} = \frac{2}{3}$, y $c_1^{(1)} = \frac{1}{3}$, por lo que

$$N_1^{(1)}(x, y) = \frac{2}{3}x + \frac{1}{3}y.$$

De manera similar, la función lineal $N_1^{(2)}(x, y)$ toma el valor 1 en $(1, 1)$ y el valor 0 tanto en $(0, 0)$ como en $(1, 0)$ satisface

$$a_1^{(2)} + b_1^{(2)}(1) + c_1^{(2)}(1) = 1,$$
$$a_1^{(2)} + b_1^{(2)}(0) + c_1^{(2)}(0) = 0,$$

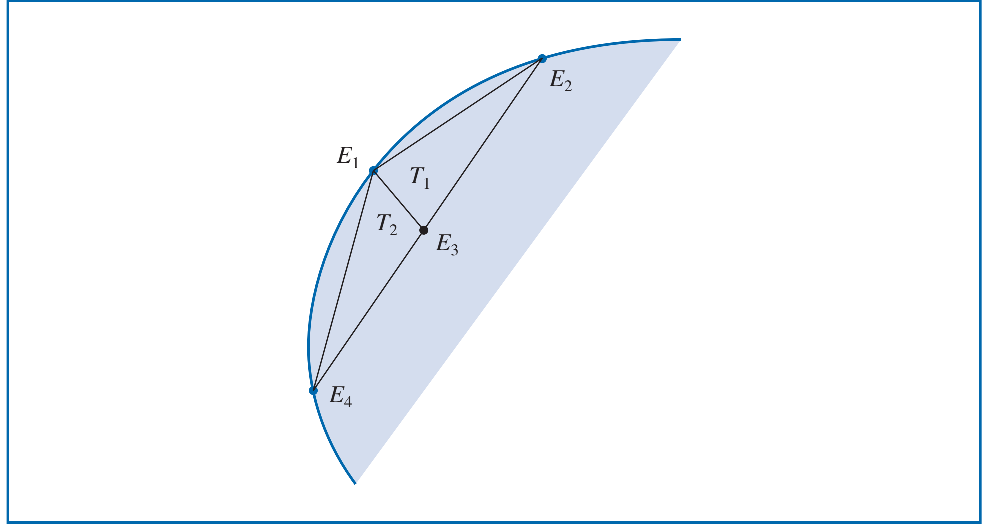
y

$$a_1^{(2)} + b_1^{(2)}(1) + c_1^{(2)}(0) = 0.$$

Esto implica que $a_1^{(2)} = 0$, $b_1^{(2)} = 0$, y $c_1^{(2)} = 1$. Como consecuencia, $N_1^{(2)}(x, y) = y$. Note que $N_1^{(1)}(x, y) = N_1^{(2)}(x, y)$ sobre la frontera común de T_1 y T_2 puesto que $y = x$. ■

Considere la figura 12.16, la parte izquierda superior de la región mostrada en la figura 12.12. Nosotros generaremos las entradas en la matriz A que corresponde a los nodos mostrados en esta figura.

Figura 12.16



Para simplicidad, suponemos que E_1 es uno de los nodos en $\mathcal{S}_1$, donde la condición de frontera $u(x, y) = g(x, y)$ es impuesta. La relación entre los nodos y los vértices de los triángulos para esta parte es

$$E_1 = V_3^{(1)} = V_1^{(2)}, \; E_4 = V_2^{(2)}, \; E_3 = V_2^{(1)} = V_3^{(2)}, \quad \text{y} \quad E_2 = V_1^{(1)}.$$

Puesto que ϕ_1 y ϕ_3 son ambos diferentes a cero en T_1 y T_2, las entradas $\alpha_{1,3} = \alpha_{3,1}$ se calculan por medio de

$$\begin{aligned}\alpha_{1,3} &= \iint_D \left[p \frac{\partial \phi_1}{\partial x} \frac{\partial \phi_3}{\partial x} + q \frac{\partial \phi_1}{\partial y} \frac{\partial \phi_3}{\partial y} - r\phi_1\phi_3 \right] dx\, dy \\ &= \iint_{T_1} \left[p \frac{\partial \phi_1}{\partial x} \frac{\partial \phi_3}{\partial x} + q \frac{\partial \phi_1}{\partial y} \frac{\partial \phi_3}{\partial y} - r\phi_1\phi_3 \right] dx\, dy \\ &\quad + \iint_{T_2} \left[p \frac{\partial \phi_1}{\partial x} \frac{\partial \phi_3}{\partial x} + q \frac{\partial \phi_1}{\partial y} \frac{\partial \phi_3}{\partial y} - r\phi_1\phi_3 \right] dx\, dy.\end{aligned}$$

En el triángulo T_1,

$$\phi_1(x, y) = N_3^{(1)}(x, y) = a_3^{(1)} + b_3^{(1)}x + c_3^{(1)}y$$

y

$$\phi_3(x, y) = N_2^{(1)}(x, y) = a_2^{(1)} + b_2^{(1)}x + c_2^{(1)}y,$$

por lo que para todas las (x, y),

$$\frac{\partial \phi_1}{\partial x} = b_3^{(1)}, \quad \frac{\partial \phi_1}{\partial y} = c_3^{(1)}, \quad \frac{\partial \phi_3}{\partial x} = b_2^{(1)}, \quad \text{y} \quad \frac{\partial \phi_3}{\partial y} = c_2^{(1)}.$$

De igual forma, en T_2,

$$\phi_1(x, y) = N_1^{(2)}(x, y) = a_1^{(2)} + b_1^{(2)}x + c_1^{(2)}y$$

y

$$\phi_3(x, y) = N_3^{(2)}(x, y) = a_3^{(2)} + b_3^{(2)}x + c_3^{(2)}y,$$

por lo que todas las (x, y),

$$\frac{\partial \phi_1}{\partial x} = b_1^{(2)}, \quad \frac{\partial \phi_1}{\partial y} = c_1^{(2)}, \quad \frac{\partial \phi_3}{\partial x} = b_3^{(2)}, \quad \text{y} \quad \frac{\partial \phi_3}{\partial y} = c_3^{(2)}.$$

Por lo tanto,

$$\begin{aligned}\alpha_{1,3} = b_3^{(1)} b_2^{(1)} \iint_{T_1} p\, dx\, dy + c_3^{(1)} c_2^{(1)} \iint_{T_1} q\, dx\, dy \\ - \iint_{T_1} r\left(a_3^{(1)} + b_3^{(1)}x + c_3^{(1)}y\right)\left(a_2^{(1)} + b_2^{(1)}x + c_2^{(1)}y\right) dx\, dy \\ + b_1^{(2)} b_3^{(2)} \iint_{T_2} p\, dx\, dy + c_1^{(2)} c_3^{(2)} \iint_{T_2} q\, dx\, dy \\ - \iint_{T_2} r\left(a_1^{(2)} + b_1^{(2)}x + c_1^{(2)}y\right)\left(a_3^{(2)} + b_3^{(2)}x + c_3^{(2)}y\right) dx\, dy.\end{aligned}$$

Todas las integrales dobles sobre D reducen las dobles integrales sobre los triángulos. El procedimiento normal es calcular todas las integrales posibles sobre los triángulos y acumularlas en la entrada correcta α_{ij} en A. De igual forma, las dobles integrales de la forma

$$\iint_D f(x, y)\phi_i(x, y)\, dx\, dy$$

se calculan sobre triángulos y, después, se acumulan en la entrada correcta β_i del vector **b**. Por ejemplo, para determinar β_1, necesitamos

$$\begin{aligned}-\iint_D f(x, y)\phi_1(x, y)\, dx\, dy = -\iint_{T_1} f(x, y)\left[a_3^{(1)} + b_3^{(1)}x + c_3^{(1)}y\right] dx\, dy \\ - \iint_{T_2} f(x, y)\left[a_1^{(2)} + b_1^{(2)}x + c_1^{(2)}y\right] dx\, dy.\end{aligned}$$

Puesto que E_1 es un vértice tanto de T_1 como de T_2, la parte de β_1 está contribuida por ϕ_1, restringida a T_1 y el resto por ϕ_1 restringida por T_2. Además, los nodos que se encuentran en $\mathcal{S}_2$ tienen incorporadas integrales lineales a sus entradas en A y **b**.

El algoritmo 12.5 realiza el método de elementos finitos en una ecuación diferencial elíptica de segundo orden. El algoritmo establece inicialmente todos los valores de la matriz A y el vector **b** en 0 y, después, realiza todas las integraciones en todos los triángulos, añade estos valores a las entradas adecuadas en A y **b**.

ALGORITMO
12.5

Método de elementos finitos

Para aproximar la solución de la ecuación diferencial parcial

$$\frac{\partial}{\partial x}\left(p(x, y)\frac{\partial u}{\partial x}\right) + \frac{\partial}{\partial y}\left(q(x, y)\frac{\partial u}{\partial y}\right) + r(x, y)u = f(x, y), \quad (x, y) \in D$$

sujeta a las condiciones de frontera

$$u(x, y) = g(x, y), \quad (x, y) \in \mathcal{S}_1$$

y

$$p(x, y)\frac{\partial u}{\partial x}(x, y)\cos\theta_1 + q(x, y)\frac{\partial u}{\partial y}(x, y)\cos\theta_2 + g_1(x, y)u(x, y) = g_2(x, y),$$

$$(x, y) \in \mathcal{S}_2,$$

donde $\mathcal{S}_1 \cup \mathcal{S}_2$ es la frontera de D y θ_1 y θ_2 son los ángulos de dirección de la normal para la frontera:

Paso 0 Divida la región D en triángulos $T_1, \ldots, T_M$ de tal forma que:
$T_1, \ldots, T_K$ son los triángulos sin bordes en $\mathcal{S}_1$ o $\mathcal{S}_2$;
(*Nota:* $K = 0$ *implica que ningún triángulo es interior para* D.)
$T_{K+1}, \ldots, T_N$ son triángulos con al menos un borde en $\mathcal{S}_2$;
$T_{N+1}, \ldots, T_M$ son los triángulos restantes.
(*Nota:* $M = N$ *implica que todos los triángulos tienen bordes en* $\mathcal{S}_2$.)
Etiquete los tres vértices del triángulo T_i mediante
$\left(x_1^{(i)}, y_1^{(i)}\right), \left(x_2^{(i)}, y_2^{(i)}\right)$, y $\left(x_3^{(i)}, y_3^{(i)}\right)$.
Etiquete los nodos (vertices) $E_1, \ldots, E_m$ donde
$E_1, \ldots, E_n$ están en $D \cup \mathcal{S}_2$ y $E_{n+1}, \ldots, E_m$ están en $\mathcal{S}_1$.
(*Nota:* $n = m$ *implica que* $\mathcal{S}_1$ *no contiene nodos*.)

ENTRADA enteros K, N, M, n, m; vértices $\left(x_1^{(i)}, y_1^{(i)}\right), \left(x_2^{(i)}, y_2^{(i)}\right), \left(x_3^{(i)}, y_3^{(i)}\right)$ para cada $i = 1, \ldots, M$; nodos E_j para cada $j = 1, \ldots, m$.

(*Nota: Todo lo que se necesita es un medio para corresponder a un vértice* $\left(x_k^{(i)}, y_k^{(i)}\right)$ *para un nodo* $E_j = (x_j, y_j)$.)

SALIDA constantes $\gamma_1, \ldots, \gamma_m$; $a_j^{(i)}, b_j^{(i)}, c_j^{(i)}$ para cada $j = 1, 2$ y $i = 1, \ldots, M$.

Paso 1 Para $l = n+1, \ldots, m$ determine $\gamma_l = g(x_l, y_l)$. (*Nota:* $E_l = (x_l, y_l)$.)

Paso 2 Para $i = 1, \ldots, n$
determine $\beta_i = 0$;
para $j = 1, \ldots, n$ determine $\alpha_{i,j} = 0$.

Paso 3 Para $i = 1, \ldots, M$

$$\text{determine } \Delta_i = \det \begin{vmatrix} 1 & x_1^{(i)} & y_1^{(i)} \\ 1 & x_2^{(i)} & y_2^{(i)} \\ 1 & x_3^{(i)} & y_3^{(i)} \end{vmatrix};$$

$$a_1^{(i)} = \frac{x_2^{(i)} y_3^{(i)} - y_2^{(i)} x_3^{(i)}}{\Delta_i}; \qquad b_1^{(i)} = \frac{y_2^{(i)} - y_3^{(i)}}{\Delta_i}; \qquad c_1^{(i)} = \frac{x_3^{(i)} - x_2^{(i)}}{\Delta_i};$$

$$a_2^{(i)} = \frac{x_3^{(i)} y_1^{(i)} - y_3^{(i)} x_1^{(i)}}{\Delta_i}; \qquad b_2^{(i)} = \frac{y_3^{(i)} - y_1^{(i)}}{\Delta_i}; \qquad c_2^{(i)} = \frac{x_1^{(i)} - x_3^{(i)}}{\Delta_i};$$

$$a_3^{(i)} = \frac{x_1^{(i)} y_2^{(i)} - y_1^{(i)} x_2^{(i)}}{\Delta_i}; \qquad b_3^{(i)} = \frac{y_1^{(i)} - y_2^{(i)}}{\Delta_i}; \qquad c_3^{(i)} = \frac{x_2^{(i)} - x_1^{(i)}}{\Delta_i};$$

para $j = 1, 2, 3$
defina $N_j^{(i)}(x, y) = a_j^{(i)} + b_j^{(i)} x + c_j^{(i)} y$.

Paso 4 Para $i = 1, \ldots, M$ (*Las integrales en los pasos* 4 *y* 5 *se pueden evaluar por medio de integración numérica*.)
para $j = 1, 2, 3$
para $k = 1, \ldots, j$ (*Calcule todas las integrales dobles sobre los triángulos*.)

$$\text{determine } z_{j,k}^{(i)} = b_j^{(i)} b_k^{(i)} \iint_{T_i} p(x, y)\, dx\, dy + c_j^{(i)} c_k^{(i)} \iint_{T_i} q(x, y)\, dx\, dy$$
$$- \iint_{T_i} r(x, y) N_j^{(i)}(x, y) N_k^{(i)}(x, y)\, dx\, dy;$$

$$\text{determine } H_j^{(i)} = -\iint_{T_i} f(x, y) N_j^{(i)}(x, y)\, dx\, dy.$$

Paso 5 Para $i = K+1, \ldots, N$ (*Calcule todas las integrales lineales*.)
para $j = 1, 2, 3$
para $k = 1, \ldots, j$

$$\text{determine } J_{j,k}^{(i)} = \int_{\mathcal{S}_2} g_1(x, y) N_j^{(i)}(x, y) N_k^{(i)}(x, y)\, dS;$$

$$\text{determine } I_j^{(i)} = \int_{\mathcal{S}_2} g_2(x, y) N_j^{(i)}(x, y)\, dS.$$

Paso 6 Para $i = 1, \ldots, M$ haga los pasos 7–12. (*Ensamble las integrales sobre cada triángulo en el sistema lineal*.)

Paso 7 Para $k = 1, 2, 3$ haga los pasos 8–12.

Paso 8 Encuentre l de modo que $E_l = \left(x_k^{(i)}, y_k^{(i)}\right)$.

Paso 9 Si $k > 1$ entonces para $j = 1, \ldots, k-1$ haga los pasos 10, 11.

Paso 10 Encuentre t de modo que $E_t = \left(x_j^{(i)}, y_j^{(i)}\right)$.

Paso 11 Si $l \leq n$ determine

si $t \leq n$ entonces determine $\alpha_{lt} = \alpha_{lt} + z_{k,j}^{(i)}$;

$\alpha_{tl} = \alpha_{tl} + z_{k,j}^{(i)}$

si no determine $\beta_l = \beta_l - \gamma_t z_{k,j}^{(i)}$

si no

si $t \leq n$ entonces determine $\beta_t = \beta_t - \gamma_l z_{k,j}^{(i)}$.

Paso 12 Si $l \leq n$ entonces determine $a_{ll} = \alpha_{ll} + z_{k,k}^{(i)}$;

$\beta_l = \beta_l + H_k^{(i)}$.

Paso 13 Para $i = K+1, \ldots, N$ haga los pasos 14–19. (*Ensamble las integrales lineales en el sistema lineal*.)

Paso 14 Para $k = 1, 2, 3$ haga los pasos 15–19.

Paso 15 Encuentre l de modo que $E_l = \left(x_k^{(i)}, y_k^{(i)}\right)$.

Paso 16 Si $k > 1$ entonces para $j = 1, \ldots, k-1$ haga los pasos 17, 18.

Paso 17 Encuentre t de modo que $E_t = \left(x_j^{(i)}, y_j^{(i)}\right)$.

Paso 18 Si $l \leq n$ entonces

si $t \leq n$ entonces determine $\alpha_{lt} = \alpha_{lt} + J_{k,j}^{(i)}$;

$\alpha_{tl} = \alpha_{tl} + J_{k,j}^{(i)}$

si no determine $\beta_l = \beta_l - \gamma_t J_{k,j}^{(i)}$

si no

si $t \leq n$ entonces determine $\beta_t = \beta_t - \gamma_l J_{k,j}^{(i)}$.

Paso 19 Si $l \leq n$ entonces determine $\alpha_{ll} = \alpha_{ll} + J_{k,k}^{(i)}$;

$\beta_l = \beta_l + I_k^{(i)}$.

Paso 20 Resolver el sistema lineal $A\mathbf{c} = \mathbf{b}$ donde $A = (\alpha_{l,t})$, $\mathbf{b} = (\beta_l)$ y $\mathbf{c} = (\gamma_t)$ para $1 \leq l \leq n$ y $1 \leq t \leq n$.

Paso 21 SALIDA $(\gamma_1, \ldots, \gamma_m)$.
(*Para cada* $k = 1, \ldots, m$ *sea* $\phi_k = N_j^{(i)}$ *en* T_i *si* $E_k = \left(x_j^{(i)}, y_j^{(i)}\right)$.
Entonces $\phi(x, y) = \sum_{k=1}^{m} \gamma_k \phi_k(x, y)$ *aproxima* $u(x, y)$ *en* $D \cup \mathcal{S}_1 \cup \mathcal{S}_2$.)

Paso 22 Para $i = 1, \ldots, M$

para $j = 1, 2, 3$ SALIDA $\left(a_j^{(i)}, b_j^{(i)}, c_j^{(i)}\right)$.

Paso 23 PARE. (*El procedimiento está completo*.) ■

Ilustración La temperatura $u(x, y)$, en una región bidimensional D satisface la ecuación de Laplace

$$\frac{\partial^2 u}{\partial x^2}(x, y) + \frac{\partial^2 u}{\partial y^2}(x, y) = 0 \qquad \text{en } D.$$

Considere la región D mostrada en la figura 12.17 con condiciones en la frontera determinadas por

$$u(x, y) = 4, \qquad \text{para } (x, y) \in L_6 \text{ y } (x, y) \in L_7;$$

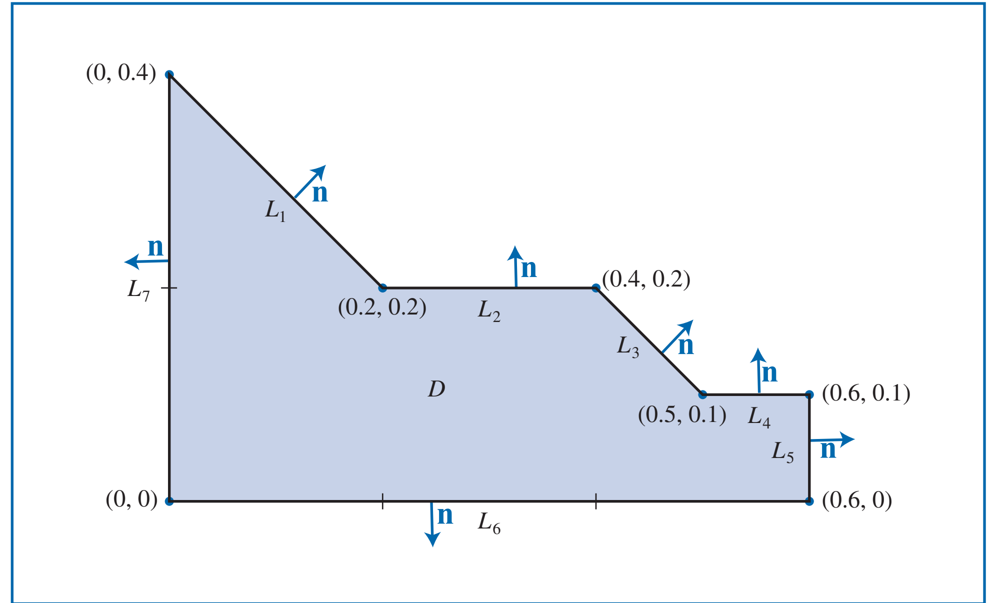
$$\frac{\partial u}{\partial \mathbf{n}}(x, y) = x, \qquad \text{para } (x, y) \in L_2 \text{ y } (x, y) \in L_4;$$

$$\frac{\partial u}{\partial \mathbf{n}}(x, y) = y, \qquad \text{para } (x, y) \in L_5;$$

$$\frac{\partial u}{\partial \mathbf{n}}(x, y) = \frac{x + y}{\sqrt{2}}, \qquad \text{para } (x, y) \in L_1 \text{ y } (x, y) \in L_3,$$

donde $\partial u/\partial \mathbf{n}$ denota la derivada direccional en la dirección de la normal $\mathbf{n}$ para la frontera de la región D en el punto (x, y).

Figura 12.17

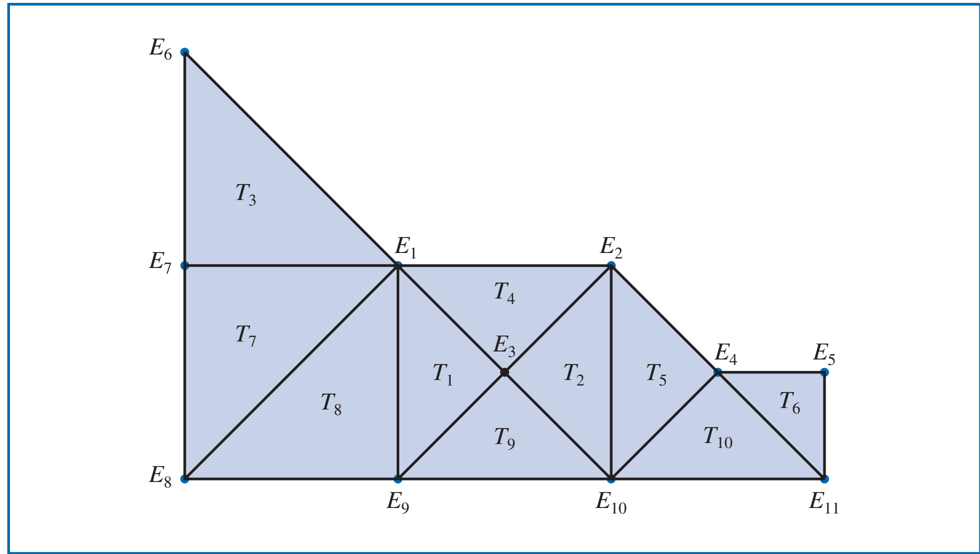


Primero subdividimos D en triángulos con la etiqueta sugerida en el paso 0 del algoritmo. Para este ejemplo, $\mathcal{S}_1 = L_6 \cup L_7$ y $\mathcal{S}_2 = L_1 \cup L_2 \cup L_3 \cup L_4 \cup L_5$. El etiquetado de los triángulos se muestra en la figura 12.18.

La condición en la frontera $u(x, y)$ sobre L_6 y L_7 implica que $\gamma_t = 4$ cuando $t = 6, 7, \ldots, 11$, es decir, en los nodos $E_6, E_7, \ldots, E_{11}$. Para determinar los valores de γ_l para $l = 1, 2, \ldots, 5$, aplique los pasos restantes del algoritmo y genere la matriz

$$A = \begin{bmatrix} 2.5 & 0 & -1 & 0 & 0 \\ 0 & 1.5 & -1 & -0.5 & 0 \\ -1 & -1 & 4 & 0 & 0 \\ 0 & -0.5 & 0 & 2.5 & -0.5 \\ 0 & 0 & 0 & -0.5 & 1 \end{bmatrix}$$

Figura 12.18



y el vector

$$\mathbf{b} = \begin{bmatrix} 6.066\bar{6} \\ 0.063\bar{3} \\ 8.0000 \\ 6.056\bar{6} \\ 2.031\bar{6} \end{bmatrix}.$$

La solución de la ecuación $A\mathbf{c} = \mathbf{b}$ es

$$\mathbf{c} = \begin{bmatrix} \gamma_1 \\ \gamma_2 \\ \gamma_3 \\ \gamma_4 \\ \gamma_5 \end{bmatrix} = \begin{bmatrix} 4.0383 \\ 4.0782 \\ 4.0291 \\ 4.0496 \\ 4.0565 \end{bmatrix}.$$

Resolver este sistema da la siguiente aproximación para la solución de la ecuación de Laplace y las condiciones de frontera en los triángulos respectivos:

$$T_1: \quad \phi(x, y) = 4.0383(1 - 5x + 5y) + 4.0291(-2 + 10x) + 4(2 - 5x - 5y),$$

$$T_2: \quad \phi(x, y) = 4.0782(-2 + 5x + 5y) + 4.0291(4 - 10x) + 4(-1 + 5x - 5y),$$

$$T_3: \quad \phi(x, y) = 4(-1 + 5y) + 4(2 - 5x - 5y) + 4.0383(5x),$$

$$T_4: \quad \phi(x, y) = 4.0383(1 - 5x + 5y) + 4.0782(-2 + 5x + 5y) + 4.0291(2 - 10y),$$

$$T_5: \quad \phi(x, y) = 4.0782(2 - 5x + 5y) + 4.0496(-4 + 10x) + 4(3 - 5x - 5y),$$

$$T_6: \quad \phi(x, y) = 4.0496(6 - 10x) + 4.0565(-6 + 10x + 10y) + 4(1 - 10y),$$

$$T_7: \quad \phi(x, y) = 4(-5x + 5y) + 4.0383(5x) + 4(1 - 5y),$$

$$T_8: \quad \phi(x, y) = 4.0383(5y) + 4(1 - 5x) + 4(5x - 5y),$$

$$T_9: \quad \phi(x, y) = 4.0291(10y) + 4(2 - 5x - 5y) + 4(-1 + 5x - 5y),$$

$$T_{10}: \quad \phi(x, y) = 4.0496(10y) + 4(3 - 5x - 5y) + 4(-2 + 5x - 5y).$$

La solución real del problema de valor en la frontera es $u(x, y) = xy + 4$. La tabla 12.7 compara el valor de u con el valor de ϕ en E_i, para cada $i = 1, \ldots, 5$. ■

Tabla 12.7

x	y	$\phi(x, y)$	$u(x, y)$	$\lvert\phi(x, y) - u(x, y)\rvert$
0.2	0.2	4.0383	4.04	0.0017
0.4	0.2	4.0782	4.08	0.0018
0.3	0.1	4.0291	4.03	0.0009
0.5	0.1	4.0496	4.05	0.0004
0.6	0.1	4.0565	4.06	0.0035

Normalmente, el error de los problemas elípticos de segundo orden del tipo (12.27) con funciones de coeficiente suave es $O(h^2)$, donde h es el diámetro máximo de los elementos triangulares. Las funciones base bilineales por tramos en elementos rectangulares también se espera que proporcionen resultados $O(h^2)$, donde h es la longitud diagonal máxima de los elementos rectangulares. Otras clases de funciones base se pueden usar para proporcionar resultados $O(h^4)$, pero la construcción es más compleja. Los teoremas de error eficiente para los métodos de elementos finitos son difíciles de establecer y de aplicar ya que la exactitud de la aproximación depende de la regularidad de la frontera, así como de las propiedades de continuidad de la solución.

El método de elementos finitos también se puede aplicar a las ecuaciones diferenciales parciales parabólicas e hiperbólicas, pero el procedimiento de minimización es más difícil. Un buen estudio sobre las ventajas y las técnicas del método de elementos finitos aplicado a varios problemas físicos se puede encontrar en el artículo de [Fi]. Para un análisis más extenso, consulte [SF], [ZM] o [AB].

La sección Conjunto de ejercicios 12.4 está disponible en línea. Encuentre la ruta de acceso en las páginas preliminares.

12.5 Software numérico

Una de las subrutinas a partir de la Biblioteca IMSL se usa para la ecuación diferencial parcial

$$\frac{\partial u}{\partial t} = F\left(x, t, u, \frac{\partial u}{\partial x}, \frac{\partial^2 u}{\partial x^2}\right),$$

con condiciones de frontera

$$\alpha(x, t)u(x, t) + \beta(x, t)\frac{\partial u}{\partial x}(x, t) = \gamma(x, t).$$

La rutina está basada en colocación en los puntos gaussianos en el eje x para cada valor de t y usa splines cúbicos de Hermite como funciones base. Otra subrutina de IMSL se usa para resolver la ecuación de Poisson en un rectángulo. El método de solución está basado en una selección de diferencias finitas de segundo y cuarto orden en una malla uniforme.

La librería NAG tiene un número de subrutinas para ecuaciones diferenciales parciales. Se utiliza una subrutina para la ecuación de Laplace en un dominio arbitrario en el plano xy y otro para resolver una ecuación diferencial parcial parabólica mediante el método de líneas.

Existen paquetes especializados, como NASTRAN, que consisten en códigos para el método de elementos finitos. Estos paquetes son populares en aplicaciones de ingeniería. El paquete FISHPACK en la biblioteca Netlib se usa para resolver ecuaciones diferenciales parciales elípticas separables. Los códigos generales para las ecuaciones diferenciales parciales son difíciles de escribir debido al problema de especificación de dominios diferentes a figuras geométricas comunes. Actualmente, la investigación en el área de solución de ecuaciones diferenciales parciales está muy activa.

Las secciones Preguntas de análisis, Conceptos clave y Revisión del capítulo están disponibles en línea. Encuentre la ruta de acceso en las páginas preliminares.



Análisis numérico, 10a. ed., se escribió para que los estudiantes de ingeniería, matemáticas, ciencias de la computación puedan usarlo en los cursos sobre la teoría y la aplicación de técnicas de aproximación numérica.

Prácticamente todos los conceptos en el texto están ilustrados con un ejemplo y contiene más de 2500 ejercicios probados en clase que van desde aplicaciones fundamentales de métodos y algoritmos hasta generalizaciones y extensiones de la teoría. Además, los conjuntos de ejercicios incluyen varios problemas aplicados de diversas áreas de la ingeniería, así como de la física, la informática, la biología y las ciencias económicas y sociales. Las aplicaciones, seleccionadas de forma clara y concisa, demuestran la manera en la que las técnicas numéricas se aplican en situaciones de la vida real.

CENGAGE Learning®

Visite nuestro sitio en http://latinoamerica.cengage.com

